



# Prediction of extreme precipitation events: combining process-based with machine learning models

Jan Philipp Heß, M.Sc.

Vollständiger Abdruck der von der TUM School of Engineering and Design  
der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. rer. nat. Martin Werner  
Prüfer\*innen der Dissertation: 1. Prof. Dr. rer. nat. Niklas Boers  
2. Prof. Dr. rer. nat. habil. Marco Körner  
3. Assoc. Prof. Dr. Nikki Vercauteren

Die Dissertation wurde am 06.02.2023 bei der Technischen Universität München  
eingereicht und durch die TUM School of Engineering and Design am 07.07.2023  
angenommen.

## **Copyright**

Prediction of extreme precipitation events: combining process-based with machine learning models

©Philipp Hess, Technical University of Munich, Munich, Germany

# Acknowledgements

This work would not have been possible without the support of my mentors, colleagues, friends and family over the past three years. Although it is difficult to put my gratefulness into a few words, this is an attempt.

I am deeply thankful to my supervisor Dr Niklas Boers, for his constant support, who always encouraged me and helped me to keep confidence in my work during challenging times. I learned so much from our extensive discussions about the project's research and beyond. Thanks to his flexibility, it was possible to continue working effectively despite the isolation during the COVID-19 pandemic. He gave me the freedom to follow my own interests and, at the same time, helped me not to lose sight of the goal.

I am grateful to Dr Marco Körner for his stimulating advice and his help as my mentor at TUM. I would like to sincerely thank Dr Markus Drüke and Dr Stefan Petri for the great collaboration, which I enjoyed very much and taught me a lot about Earth system modelling. Brainstorming ideas and working together with Felix Strnad has been a lot of fun, and I would like to thank him also for proofreading this thesis. The effective collaboration with Dr Stefan Lange was also an enrichment for which I would like to thank him.

I really enjoy the open and pleasant atmosphere in our research group during online meetings, lunch breaks or chats over a cup of coffee. Therefore, I would like to thank my colleagues (and in no particular order): Dr Takahito Mitsui, Keno Riechers, Dr Maximilian Gelbrecht, Alistair White, Lana Blaschke, Dr Sebastian Bathiany, Dr Christof Schötz, Michael Aich, Dr Da Nian, Andreas Morr, Maya Ben-Yami and Dr Qin Hu.

I would like to thank Karin Haußmann, Sophia Kostial and Gabriele Pilz for their tremendous help in all bureaucratic and organizational matters.

Further, I would like to thank my girlfriend, my brother and my friends for their encouragement and welcomed distractions from work. My parents have always supported me in my decisions, and without them, all this would not have been possible. Thank you.

This work was supported by the Volkswagen Foundation.



# List of publications

This cumulative dissertation is based on the following publications.

- P1 Hess, P., & Boers, N. (2022).** Deep Learning for Improving Numerical Weather Prediction of Heavy Rainfall. *Journal of Advances in Modeling Earth Systems*, 14(3), e2021MS002765.  
DOI: <https://doi.org/10.1029/2021MS002765>
- P2 Hess, P., Drüke, M., Petri, S., Strnad, F. M., & Boers, N. (2022).** Physically constrained generative adversarial networks for improving precipitation fields from Earth system models. *Nature Machine Intelligence*, 4(10), 828-839.  
DOI: <https://doi.org/10.1038/s42256-022-00540-1>
- P3 Hess, P., Lange, S., & Boers, N. (2022).** Deep Learning for bias-correcting comprehensive high-resolution Earth system models. Submitted to *Proceedings of the National Academy of Sciences*  
DOI (arXiv preprint): <https://doi.org/10.48550/arXiv.2301.01253>



# Summary

## Background

Modelling the Earth system and its components, such as the atmosphere, oceans, biosphere, or ice sheets, is a daunting task. It includes many physical processes that interact across wide ranges of spatial and temporal scales, i.e., from millimeters to thousands of kilometers and milliseconds to millennia.

Numerical models that integrate the fundamental physical equations of motion that govern the dynamics, as well as thermodynamics, are our primary tool to forecast tomorrow's weather or to project climate scenarios for the coming decades. While significant improvements to the models have been made over the past decades, substantial errors and uncertainties remain (Bauer et al., 2015; Schneider et al., 2017b). Due to computational constraints, the discretized simulations have a limited spatial and temporal resolution, and important physical processes remain unresolved. Small-scale processes that cannot be fully resolved have to be approximated as parameterizations, i.e., written as functions of the resolved variables.

Precipitation, formed by condensation of water vapour in the atmosphere, is the result of an interaction of processes on large ranges of scales, i.e., from microphysical droplet interactions over atmospheric turbulence to frontal weather systems. Hence, the discretized nature of numerical models strongly affects the representation of precipitation and can lead to biases in the simulations with an under- or overestimation of extremes (Déqué, 2007; Cannon et al., 2015).

This is problematic, as precipitation is arguably one of the most important atmospheric variables with significant ecological and socio-economic impacts (Kotz et al., 2022), e.g., on transportation, air traffic, utility sector, agriculture, and the natural biosphere. Moreover, with anthropogenic global warming projected to increase in the coming decades, precipitation extremes are likely to increase globally in frequency and intensity (Wilcox and Donner, 2007; Fischer and Knutti, 2016). On the other hand, local trends have been found to show heterogeneous changes over the next decades (Ali et al., 2018; Traxl et al., 2021). Therefore, accurately modelling extreme precipitation is an urgent problem and will likely gain importance in the future (IPCC, 2021).

Statistical post-processing methods have therefore been developed to adjust systematic errors in the numerical simulations (Wilks, 2006; Gudmundsson et al., 2012; Cannon et al., 2015). These methods typically adjust model biases locally for each grid cell individually without taking into account spatial correlations.

In recent years, advances in computational hardware such as graphics and tensor processing units (GPUs and TPUs) (Wang et al., 2020), as well as in software libraries and algorithms, have led to a renaissance in machine

learning, and in particular, deep neural networks (DNNs) that can use spatial context for their predictions (LeCun et al., 2015).

The volume and quality of data that data-driven methods can leverage have significantly increased through the development of high-resolution remote sensing products and increasingly comprehensive simulations.

Thus, the emergence of efficient hardware and algorithms, together with large amounts of spatial-temporal training data, make deep learning methods a promising tool for correcting biases in numerical precipitation simulations.

## Scope of the dissertation

This cumulative thesis explores deep learning methods for post-processing tasks of numerical weather and climate simulations of precipitation, and particularly its extremes. More specifically, computer vision and image processing methods are investigated to improve essential precipitation characteristics that are missing in the numerical simulations. In summary, the following key research questions and challenges have been addressed in this work:

**Q1** *Can DNNs improve spatial patterns of precipitation, and in particular, the characteristic small-scale variability?*

**Challenge** Due to the relatively coarse resolution and a limited number of physical processes in numerical weather and climate models, precipitation dynamics cannot be completely resolved. The output fields in numerical simulations often appear overly smooth, i.e., spatial patterns lack characteristic small-scale variability.

**Approach** In short-term numerical weather prediction (NWP), the output of the weather model can be *directly* compared to observations for each time step. Publication P1 (chapter 2) studies the application of *supervised* convolutional neural networks (CNNs) that can use multi-scale spatial context to improve the fidelity of the simulation in a post-processing step by using suitably designed loss functions. Climate simulations do not follow observations directly, i.e., they are built to produce realistic temporal distributions over long periods. Therefore Publication P2 and P3 (chapter 3) follow an *unsupervised* approach using generative adversarial networks (GANs), where a second network becomes part of the loss function that enables realistic output fields.

**Q2** *Can DNNs accurately learn to correct the frequency distribution of precipitation extremes from biased model simulations?*

**Challenge** Modelling precipitation extremes accurately is challenging, given the strongly skewed distribution and, by definition, rare occurrence. Numerical simulations often exhibit biased distributions that over- or underestimate frequencies of extreme events.



**Approach** Publication P1 investigates this problem from a weather prediction perspective. Different loss functions and their effect on the CNN to accurately predict extremes in the upper distribution tail are compared.

Publications P2 and P3 pursue this question from a climate perspective by applying generative models that learn a transformation from a given distribution to another target distribution.

**Q3** *Can DNNs increase the forecast skill for rare precipitation events in numerical weather predictions?*

**Challenge** The main challenge here is to use the relatively sparse training data of extreme events to improve the forecast skill of the NWP predictions. More specifically, to increase the accuracy of extreme precipitation event predictions for a given time and place.

**Approach** The research leading to publication P1 also explored different input features (or predictors) that can be linked to heavy precipitation events, neural network architectures that are suitable for the multi-scale nature of spatial precipitation fields, and their effect on the DNN's forecast skill.

**Q4** *Can DNNs improve numerical climate simulations of precipitation?*

**Challenge** Climate simulations do not follow observations, as the chaotic nature of the atmosphere leads to diverging trajectories. Hence, post-processing methods have to be able to deal with *unpaired* training samples. Further challenging is the lack of ground truth for future projections and the inherent non-stationarity of the Earth's climate. This leads to out-of-sample predictions and possible violations of physical conservation laws.

**Approach** Deep learning methods for image-to-image translation are investigated in publications P2 and P3, particularly cycle-consistent generative adversarial networks (CycleGANs) (Zhu et al., 2017) that change the "style" of an image but not its overall content. Moreover, their ability to improve summary statistics over long periods (e.g. decades) by transforming single fields on short time scales (e.g. days), as well as the incorporation of physical constraints, are studied.

## Main outcomes

The main outcomes and contributions of this dissertation are summarized in the following.

**Q1 | Improved spatial patterns of precipitation** In the weather prediction context of P1, the U-Net-based CNN can strongly improve the spatial patterns of rainfall forecasts from the Integrated forecast system (IFS) (European Centre for Medium-Range Weather Forecasts, 2012) by training on data from

the satellite-based Tropical Rainfall Measurement Mission (TRMM) (Huffman et al., 2007) product. Combining a suitable architecture with a loss function incorporating a multi-scale structural similarity index measure (MS-SSIM) (Wang et al., 2003) leads to improved spatial patterns. This is evaluated with a complex wavelet-based extension of the SSIM that is invariant to small rotations and translations, i.e., evaluates the similarity of the patterns without penalizing small deviations in their position.

The CycleGAN in P2 and P3 can translate the overly smooth and blurry climate simulations into much sharper fields, s.t. the characteristic variability in spatial patterns is visually indistinguishable from the observation-based reanalysis ground truth. We evaluate the spatial fields using power spectral densities and the fractal dimension that captures the characteristics of the small-scale variability. It is shown that the developed method strongly outperforms established quantile mapping-based techniques.

**Q2 | Learning the distribution of precipitation extremes** To learn the strongly skewed distribution of precipitation, a weighted mean squared error is shown in P1 to enable the CNN to correct the frequency distributions over all precipitation values far into the upper distribution tail. A comparable skill to quantile mapping - a method that is specifically designed to correct distributions - is achieved by the CNN.

Training a CNN as a generative adversarial network in P2 and P3 produces distributions that closely match the ground truth, including the tails. It thereby strongly improves the numerical simulations without the need to engineer a suitable loss function manually. It achieves comparable or better results than the quantile mapping baseline on low-resolution simulations. Using comprehensive high-resolution Earth system simulations (GFDL-ESM4) (Krasting et al., 2018), our method performs comparably to a state-of-the-art bias correction framework (ISIMIP3BASD) (Lange, 2019).

**Q3 | Increased forecast skill of rare rainfall events** The forecast skill for (nearly) global rainfall is significantly improved by the CNN-based post-processing in P1. It uses the vertical velocity of wind that can be linked to updrafts and convection as additional input features. The architecture and loss function are suitably chosen for the multi-scale nature of precipitation patterns. The method improves continuous evaluation metrics and categorical skill scores of extreme events, outperforming several baselines.

**Q4 | Realistic and efficient climate simulations of precipitation** The CycleGAN applied in P2 and P3 can be trained naturally on the unpaired training samples from climate simulations and the observation-based ground truth. Taking the results from Q1 and Q2 together, the model can correct the simulations on short and long time scales, i.e., with respect to spatial patterns of daily fields and temporal distributions over a decade. The method is thus able to make climate simulations much more realistic. To generalise

to warmer climates unseen during the training on historical observations, a physical constraint is introduced that preserves the global precipitation sum per time step. It ensures consistency with the large-scale hydrological cycle in the simulation and enables the GAN to capture trends in global mean precipitation as expected from thermodynamic considerations. Finally, a gradient-based interpretability method is used as a sensitivity analysis of the discriminator network. It shows that the network uses the geographical regions with the largest model bias to distinguish between generated and ground truth precipitation fields.

## Conclusion and outlook

This thesis investigates the application of deep convolutional neural networks and training techniques to develop post-processing methods for weather and climate simulations. A central focus is on improving extreme precipitation events that can have a high impact.

In study P1, the CNN has shown great potential to predict extreme rainfall events that are not accurately represented in the input. Given these encouraging findings, possible extensions are probabilistic ensemble forecasting and downscaling. This would go in hand with the work on generative adversarial methods in this thesis and other recent studies (e.g. [Ravuri et al. \(2021\)](#); [Harris et al. \(2022\)](#); [Price and Rasp \(2022\)](#)).

The second part of this thesis has shown that unpaired image-to-image translation in deep learning can be used for ESM bias correction. It allows for improving the simulations in a new dimension also spatially, which is not possible with established methods. Adding further variables in future work besides precipitation should be straightforward and would enable physically consistent multivariate bias correction. Combinations with downscaling to increase spatial resolution are another promising direction for future research.

With continuing advances in deep learning algorithms, specialized hardware, and the increasing volume and quality of geospatial data, the integration of deep learning-based approaches in Earth system modelling is likely to gain importance for the foreseeable future with exciting new possibilities to advance the integration of both domains.



# Zusammenfassung

Die Modellierung des Erdsystems und seiner Komponenten, wie die Atmosphäre, Ozeane, Biosphäre oder Eisschilde, ist eine gewaltige Aufgabe. Es umfasst viele physikalische Prozesse, die in einem weiten Bereich von räumlichen und zeitlichen Skalen interagieren, von Millimetern bis zu Tausenden von Kilometern und von Millisekunden bis zu Jahrtausenden.

Numerische Modelle, die die grundlegenden physikalischen Gleichungen der Dynamik sowie der Thermodynamik integrieren, sind unser wichtigstes Instrument für die Vorhersage des Wetters von morgen oder für die Projektion von Klimaszenarien für die kommenden Jahrzehnte. Obwohl die numerischen Modelle in den letzten Jahrzehnten erheblich verbessert wurden, zeigen sie nach wie vor erhebliche Fehler und Unsicherheiten (Bauer et al., 2015; Schneider et al., 2017b). Aufgrund von beschränkten Rechenkapazitäten haben die diskretisierten Simulationen eine begrenzte räumliche und zeitliche Auflösung. Wichtige physikalische Prozesse können daher nicht vollständig aufgelöst werden. Kleinskalige Prozesse müssen deshalb als Parametrisierungen approximiert werden, d.h., als Funktionen der aufgelösten Variablen beschrieben werden.

Niederschlag, der durch die Kondensation von wasserdampfhaltiger Luft in der Atmosphäre entsteht, ist das Ergebnis einer Interaktion von Prozessen auf einem großen Bereich von Größenskalen, die von mikrophysikalischen Tröpfcheninteraktionen über atmosphärische Turbulenzen bis hin zu Fronten in Wettersystemen reichen. Daher wirkt sich die Diskretisierung der numerischen Modelle stark auf die Darstellung des Niederschlags aus und führt oft zu systematischen Fehlern in den Simulationen mit einer Unter- oder Überschätzung von Extremen (Déqué, 2007; Cannon et al., 2015).

Dies ist problematisch, da der Niederschlag wohl eine der wichtigsten atmosphärischen Variablen mit erheblichen ökologischen und sozioökonomischen Auswirkungen ist (Kotz et al., 2022), z.B. auf das Transportwesen, den Luftverkehr, Versorgungsunternehmen, die Landwirtschaft und die natürliche Biosphäre. Mit der voraussichtlich zunehmenden anthropogenen globalen Erwärmung in den kommenden Jahrzehnten, wird auch die Häufigkeit und Intensität von Niederschlagsextremen weltweit zunehmen (Wilcox and Donner, 2007; Fischer and Knutti, 2016). Die lokalen Trends weisen allerdings heterogene Veränderungen auf (Ali et al., 2018; Traxl et al., 2021). Daher ist die genaue Modellierung von Extremniederschlägen eine dringende Aufgabe und wird in Zukunft wahrscheinlich noch an Bedeutung gewinnen (IPCC, 2021).

Statistische Postprocessing Methoden wurden entwickelt, um systematische Fehler in den numerischen Simulationen zu korrigieren (Wilks, 2006; Gudmundsson et al., 2012; Cannon et al., 2015). Diese Methoden korrigieren die Modellabweichungen in der Regel lokal für jede Gitterzelle einzeln, ohne

dabei räumliche Korrelationen zu berücksichtigen.

In den letzten Jahren haben Fortschritte in der Computerhardware, wie z.B. GPUs und TPUs (Wang et al., 2020) sowie Softwarebibliotheken und Algorithmen zu einer Renaissance des maschinellen Lernens geführt, insbesondere tiefer neuronaler Netze (TNN), die den räumlichen Kontext für ihre Vorhersagen nutzen können (LeCun et al., 2015).

Der Umfang und die Qualität der Daten, die von datengetriebenen Methoden genutzt werden können, haben durch die Entwicklung hochauflösender Fernerkundungsprodukte und immer umfassenderer Simulationen erheblich zugenommen.

Das Aufkommen effizienter Hardware und Algorithmen, sowie die große Menge an raum-zeitlichen Trainingsdaten machen Deep Learning Methoden zu einem vielversprechenden Werkzeug für die Korrektur von systematischen Fehlern in Niederschlagsdaten von numerischen Wetter und Klimasimulationen.

## Umfang der Dissertation

In dieser kumulativen Dissertation werden Deep Learning Methoden für das Postprocessing von numerischen Wetter- und Klimasimulationen des Niederschlags, insbesondere von Extremen, untersucht. Genauer gesagt werden Methoden der Computer Vision (dt. computerbasiertes Sehen) und der Bildverarbeitung analysiert und verwendet, um wesentliche Niederschlagseigenschaften zu verbessern, die in den numerischen Simulationen fehlen. Zusammenfassend wurden die folgenden zentralen Forschungsfragen und Problemstellungen in dieser Arbeit behandelt:

**Q1** *Können TNN die räumlichen Muster des Niederschlags und insbesondere dessen charakteristische kleinräumige Variabilität verbessern?*

**Problemstellung** Aufgrund der relativ groben Auflösung und einer begrenzten Anzahl von physikalischen Prozessen in numerischen Wetter- und Klimamodellen kann die Niederschlagsdynamik nicht vollständig aufgelöst werden. Die ausgegebenen Felder von numerischen Simulationen erscheinen oft zu glatt, d.h., den räumlichen Mustern fehlt die charakteristische kleinräumige Variabilität.

**Ansatz** Bei numerischen Wettervorhersagen auf kurzen Zeitskalen kann jede Vorhersage des Wettermodells direkt mit Beobachtungen verglichen werden. In der Veröffentlichung P1 (Kapitel 2) wird die Anwendung von *supervised* (dt. überwachten) Convolutional Neural Networks (CNNs) untersucht, um die Genauigkeit der Simulation durch ein Postprocessing zu verbessern. Besonders die Fähigkeit von CNNs einen mehrskaligen räumlichen Kontext nutzen zu können, sowie der Einsatz geeigneter Verlustfunktionen im Training des Netzwerkes stehen dabei im Mittelpunkt. Klimasimulationen, andererseits, folgen nicht direkt den Beobachtungen, d.h., sie werden entwickelt, um realistische Verteilungen über lange

Zeiträume zu simulieren. Daher verfolgen die Veröffentlichungen P2 und P3 (Kapitel 3) einen *unsupervised* (dt. unüberwachten) Ansatz, bei dem Generative Adversarial Networks (GANs) verwendet werden. Dabei wird ein zweites Netzwerk als Teil der Verlustfunktion verwendet und die Generierung realistischer Felder ermöglicht.

**Q2** *Können TNN die Häufigkeitsverteilung von Niederschlagsextremen aus verzerrten Modellsimulationen korrigieren?*

**Problemstellung** Die genaue Modellierung von Niederschlagsextremen ist eine große Herausforderung, durch die ausgeprägte Schiefe der Verteilung und die Seltenheit der Ereignisse. Numerische Simulationen weisen oft verfälschte Verteilungen auf, die die Häufigkeit von Extremereignissen über- oder unterschätzen.

**Ansatz** In der Veröffentlichung P1 wird dieses Problem im Kontext von Wettervorhersagen untersucht. Es werden verschiedene Verlustfunktionen und ihre Auswirkungen auf die Fähigkeit des CNNs die Häufigkeitsverteilung von Extremen abzubilden verglichen. Die Veröffentlichungen P2 und P3 verfolgen diese Frage aus der Klimaperspektive, indem sie generative Modelle anwenden, die eine Transformation von einer gegebenen Verteilung in eine andere Zielverteilung lernen.

**Q3** *Können TNN die Vorhersagefähigkeit für seltene Niederschlagsereignisse in numerischen Wettervorhersagen erhöhen?*

**Problemstellung** Die Hauptherausforderung hier besteht darin, die relativ spärlichen Trainingsdaten von Extremereignissen zu nutzen, um die Vorhersagefähigkeit der Wettervorhersagen zu verbessern. In anderen Worten, die Genauigkeit der Vorhersagen extremer Niederschlagsereignisse für einen bestimmten Zeitpunkt und Ort zu erhöhen.

**Ansatz** Die Forschungsarbeit, die zur Veröffentlichung P1 führte, untersuchte verschiedene Inputvariablen (Prädiktoren), die mit Starkniederschlagsereignissen in Verbindung gebracht werden können, sowie neuronale Netzwerkarchitekturen, die für die Multiskalen-Felder räumlicher Niederschlagsverteilungen geeignet sind, und deren Auswirkung auf die Vorhersagekraft von selten Extremen.

**Q4** *Können TNN numerische Klimasimulationen des Niederschlags verbessern?*

**Problemstellung** Klimasimulationen folgen nicht genau dem beobachteten Wetter, da die chaotische Natur der Atmosphäre zu abweichenden Trajektorien führt. Daher müssen Postprocessing Methoden in der Lage sein, mit *ungepaarten* Trainingssamples umzugehen. Eine weitere Herausforderung besteht darin, dass es keine wahre Vergleichsgrundlage für künftige Projektionen gibt und das Klima von Natur aus nicht-stationär ist. Dies führt zu Vorhersagen außerhalb der Trainingsverteilung wobei es möglicherweise zur Verletzung von physikalische Erhaltungssätze durch das neuronale Netzwerk kommen kann.

**Ansatz** In den Veröffentlichungen P2 und P3 werden Deep Learning Methoden für die Bildübersetzung untersucht, insbesondere cycle-consistent (dt. etwa zykluskonsistente) Generative Adversarial Networks (CycleGANs) (Zhu et al., 2017), die den “Stil” eines Bildes verändern, nicht aber den Inhalt. Darüber hinaus wird ihre Fähigkeit von GANs untersucht, zusammenfassende Statistiken über lange Zeiträume (z.B. Jahrzehnte) zu verbessern, indem einzelne Felder auf kurzen Zeitskalen (z.B. Tage) transformiert werden, sowie dessen Einhaltung physikalischer Grenzen.

## Hauptresultate

Die Hauptergebnisse und Beiträge dieser Dissertation werden im Folgenden zusammengefasst.

**Q1 | Verbesserte räumliche Niederschlagsmuster** Im Kontext der Wettervorhersage von P1 kann das U-Net-basierte CNN die räumlichen Muster der Niederschlagsvorhersagen des numerischen IFS Modells (European Centre for Medium-Range Weather Forecasts, 2012) stark verbessern, indem es auf satellitengestützte Daten der Tropical Rainfall Measurement Mission (TRMM) (Huffman et al., 2007) trainiert wird. Die Kombination einer geeigneten Netzwerkarchitektur mit einer Verlustfunktion, die ein multiskaliges und strukturelles Ähnlichkeitsindexmaß (Wang et al., 2003) enthält, führt zu verbesserten räumlichen Mustern. Dies wird mit einer Wavelet-basierten Erweiterung des Maßes ausgewertet, die invariant gegenüber kleinen Rotationen und Translationen ist, d.h., die Ähnlichkeit von Mustern bewertet, ohne kleine Abweichungen in ihrer Position zu berücksichtigen.

Das CycleGAN in P2 und P3 kann die unrealistisch glatten und unscharfen Klimasimulationen in deutlich schärfere Felder übersetzen, die visuell nicht von den beobachtungsbasierten Reanalysedaten zu unterscheiden sind. Wir evaluieren die Qualität der räumlichen Felder mit Hilfe der spektralen Leistungsdichte sowie durch die Berechnung der fraktalen Dimension, um die kleinskalige Variabilität im Raum zu erfassen. Es wird gezeigt, dass die entwickelte Methode etablierte, auf Quantile Mapping basierende Techniken deutlich übertrifft.

**Q2 | Lernen der Häufigkeitsverteilung von Niederschlagsextremen** Zum Erlernen der schiefen Niederschlagsverteilung wird in P1 ein gewichteter mittlerer quadratischer Fehler in der Verlustfunktion verwendet. Damit lernt das CNN die Häufigkeitsverteilungen über alle Niederschlagswerte bis weit in die oberen Verteilungsränder zu korrigieren. Das CNN erreicht vergleichbare Ergebnisse wie Quantile Mapping - eine Methode, die speziell für die Korrektur von Verteilungen entwickelt wurde.

Das Training eines CNN als Generative Adversarial Network in P2 und P3 führt zu Verteilungen, die denen der Beobachtungsdaten sehr nahe kommen,



einschließlich der Ränder. Dadurch werden die numerischen Simulationen stark verbessert, ohne dass eine geeignete Verlustfunktion manuell entwickelt werden muss. Bei Simulationen mit niedriger räumlicher Auflösung werden vergleichbare oder bessere Ergebnisse erzielt als mit Quantile Mapping. Bei der Verwendung umfassender, hochauflösender Erdsystem-Simulationen (basierend auf dem GFDL-ESM4 Modell) (Krasting et al., 2018) erreicht das GAN vergleichbare Ergebnisse wie eine etablierte Methode zur Korrektur von systematischen Fehlern (ISIMIP3BASD) (Lange, 2019).

### **Q3 | Erhöhte Vorhersagefähigkeit von seltenen Niederschlagsereignissen**

Die Vorhersagefähigkeit (fast) globaler Niederschlags-Simulationen wird durch das CNN-basierte Postprocessing in P1 deutlich verbessert. Es verwendet vertikale Windgeschwindigkeiten als zusätzliche Prädiktoren, die über Aufwinde und Konvektion mit dem Niederschlag in Verbindung stehen. Die Architektur und die Verlustfunktion wurden passend für die Multiskalen-Muster der Niederschlagsfelder gewählt. Die Methode verbessert sowohl die kontinuierlichen als auch die kategorischen Evaluierungsmetriken von Extremereignissen und übertrifft dabei mehrere Vergleichsmodelle.

### **Q4 | Realistische und effiziente Klimasimulationen des Niederschlags**

Das in P2 und P3 angewandte CycleGAN kann passend auf den ungepaarten Trainingsdaten von Klimasimulationen und der beobachtungsbasierten Reanalyse trainiert werden. Nimmt man die Ergebnisse aus Q1 und Q2 zusammen, kann das CycleGAN die Simulationen sowohl auf kurzen als auch auf langen Zeitskalen korrigieren, d.h., sowohl räumliche Muster täglicher Niederschlagsfelder, als auch zeitliche Verteilungen über ein Jahrzehnt. Die Methode ist somit in der Lage, Klimasimulationen wesentlich realistischer zu machen. Um die Verallgemeinerung der Vorhersagen des neuronalen Netzes für wärmere Klimaszenarien zu erleichtern, welche während des Trainings auf historischen Beobachtungen nicht gezeigt wurden, wird eine physikalische Begrenzung eingeführt, die die globale Niederschlagssumme pro Zeitschritt erhält. Dadurch wird die Konsistenz mit dem großräumigen hydrologischen Zyklus in der Simulation sichergestellt und das GAN in die Lage versetzt, die aus thermodynamischen Überlegungen zu erwarteten Trends des global gemittelten Niederschlags zu reproduzieren. Außerdem wird eine gradientenbasierte Interpretierbarkeitsmethode als Sensitivitätsanalyse des Diskriminatornetzes verwendet. Sie zeigt, dass das Netzwerk erwartungsgemäß die geografischen Regionen mit der größten Modellfehlern verwendet, um zwischen generierten und tatsächlichen Niederschlagsfeldern zu unterscheiden.

## **Schlussfolgerung und Ausblick**

Diese Arbeit untersucht die Anwendung von tiefen Convolutional Neural Networks und Trainingstechniken zur Entwicklung von Postprocessing Methoden für Wetter- und Klimasimulationen. Ein zentraler Fokus liegt dabei

auf der Verbesserung von extremen Niederschlagsereignissen, die eine große Auswirkung haben können.

In der Arbeit P1 zeigt das CNN großes Potenzial, die Vorhersage von extremen Niederschlagsereignissen zu verbessern, welche im Netzwerkinput nicht genau dargestellt werden. Angesichts dieser ermutigenden Ergebnisse sind probabilistische Ensemblevorhersagen und Downscaling-Anwendungen als Weiterentwicklung möglich. Dies würde zu den Arbeiten mit Generative Adversarial Networks (P2 und P3) in dieser Arbeit und anderen aktuellen Studien (z.B. [Ravuri et al. \(2021\)](#); [Harris et al. \(2022\)](#); [Price and Rasp \(2022\)](#)) passen.

Der zweite Teil dieser Arbeit hat gezeigt, dass die ungepaarte Bild-zu-Bild Übersetzung in Deep Learning für die Korrektur systematischer Fehler in Erdsystem-Simulationen verwendet werden kann. Sie erlaubt es, die Simulationen in einer neuen Dimension auch räumlich zu verbessern, was mit etablierten Methoden nicht möglich ist. Die Hinzunahme weiterer physikalischer Variablen neben dem Niederschlag sollte in zukünftigen Arbeiten ohne weiteres möglich sein. Dies würde eine physikalisch konsistente multivariate Fehlerkorrektur ermöglichen. Kombinationen mit Downscaling zur Erhöhung der räumlichen Auflösung sind eine vielversprechende Richtung für zukünftige Weiterentwicklungen.

Mit den kontinuierlichen Fortschritten in der Entwicklung von Deep Learning Algorithmen, leistungsfähigerer Hardware und der zunehmenden Menge und Qualität von Geodaten wird die Integration von Deep Learning-basierten Ansätzen in der Erdsystemmodellierung in absehbarer Zukunft voraussichtlich an Bedeutung gewinnen und spannende neue Möglichkeiten eröffnen.

# Contents

<b>Acknowledgements</b>	<b>iii</b>
<b>List of publications</b>	<b>v</b>
<b>Summary</b>	<b>vii</b>
<b>Zusammenfassung</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Precipitation . . . . .	2
1.1.1 Physical processes . . . . .	2
1.1.2 Statistical characteristics . . . . .	4
1.1.3 Observations . . . . .	5
1.2 Modelling weather and climate . . . . .	7
1.2.1 Numerical forecasts and projections . . . . .	8
1.2.2 Statistical post-processing . . . . .	11
1.3 Deep learning for post-processing . . . . .	15
1.3.1 Machine learning basics . . . . .	15
1.3.2 Computer vision and image processing . . . . .	20
1.3.3 Combining domain knowledge with machine learning . . . . .	24
1.4 Outline of the thesis . . . . .	27
<b>2 Post-processing rainfall forecasts with deep neural networks</b>	<b>29</b>
2.1 P1   Deep learning for improving numerical weather prediction of heavy rainfall . . . . .	30
<b>3 Generative models for improving precipitation fields from climate simulations</b>	<b>53</b>
3.1 P2   Physically constrained generative adversarial networks for improving precipitation fields from Earth system models . . . . .	55
3.2 P3   Deep Learning for bias-correcting comprehensive high-resolution Earth system models . . . . .	93
<b>4 Conclusion</b>	<b>113</b>
4.1 P1   Post-processing rainfall forecasts with deep neural networks	113
4.1.1 Main outcomes . . . . .	113
4.1.2 Context . . . . .	114
4.1.3 Outlook . . . . .	115
4.2 P2 and P3   Generative models for improving precipitation fields from Earth system simulations . . . . .	116
4.2.1 Main outcomes . . . . .	117

4.2.2	Context . . . . .	118
4.2.3	Outlook . . . . .	119
4.3	Future developments . . . . .	120
	<b>List of figures</b>	<b>123</b>
	<b>Bibliography</b>	<b>125</b>

# 1 Introduction

Accurately modelling precipitation and its extremes is important due to the high ecological and socio-economic impacts (Kotz et al., 2022). Precipitation is a fundamental component of the global water cycle, where sufficient freshwater resources are necessary to support the majority of living organisms on Earth, including plants, animals and human society. Long-term changes in precipitation can have important consequences for the planet’s biosphere and vegetation, including crop yields in agriculture. Extreme events can cause severe flooding, landslides, and infrastructure damage, while the absence of rainfall and droughts can have similarly severe effects. Transportation also depends on sufficiently high river water levels and the right conditions on streets and landing platforms. Therefore, accurate short-term predictions between a few hours to days are necessary for disaster prevention and mitigation, while reliable long-term projections are required to assess the impacts caused by anthropogenic global warming.

However, skillfully predicting precipitation, particularly its extremes, is challenging. The complex interaction of physical processes across a large range of spatial and temporal scales, the high variability, and a strongly non-Gaussian and skewed distribution must be captured closely.

Numerical models that simulate the atmosphere on discretized grids with a finite resolution and model complexity cannot resolve important processes on small scales due to computational constraints. Hence, they commonly exhibit large biases in their output and under- or overestimate extreme events (Déqué, 2007; Cannon et al., 2015). Therefore, accurate weather forecasts and realistic climate projections remain challenging despite the progress that has been made over the past decades (Bauer et al., 2015; Schneider et al., 2017b).

Statistical post-processing methods aim to correct such biases (Wilks, 2006; Gudmundsson et al., 2012; Cannon et al., 2015) but have not been developed to use a larger spatial context efficiently for this task.

In recent years three main trends have started to enable new data-driven approaches that can potentially address this short-coming: (i) the availability of new high-resolution observational datasets and comprehensive simulations with global coverage (Hersbach et al., 2020), (ii) advances in computer vision algorithms and software libraries for optimizing deep learning models with large parameters spaces (LeCun et al., 2015) and (iii) the development of specialized hardware, such as graphics and tensor processing units (GPUs and TPUs), that can efficiently process large volumes of data and train deep neural networks.

Based on these advances, the central topic of this thesis is the application of deep learning-based post-processing methods that use spatial context to improve precipitation simulations of short-term weather and long-term cli-

mate. More specifically, this thesis investigates whether deep convolutional neural networks can be used to improve spatial patterns and relative frequency distributions of precipitation simulations. In the weather prediction context, methods for improving the forecast skill of precipitation and particularly extreme events are investigated. When post-processing climate model simulations, challenges that arise from the chaotic nature of the atmosphere, and the inherent non-stationarity of the system, are also addressed.

This thesis is organized as follows. In the first part of the introduction, physical processes and their multi-scale character that lead to the generation of precipitation and the challenges involved in its measurement are discussed. The section concludes with a brief discussion of the characteristic statistical properties of precipitation.

The second part focuses on modelling aspects of the Earth system, emphasising the main differences between weather prediction and climate projections to help distinguish the different problems addressed in the publications of the thesis. It is meant to introduce the main sources of errors in numerical simulations of the atmosphere and how they can be improved with post-processing methods.

The last part introduces deep learning concepts for post-processing spatial fields, including training techniques, architectures for computer vision applications, and hybrid approaches combining physical and data-driven methods that have been applied in the studies of this thesis.

After the introduction, the publications are included and summarized in two chapters on weather forecasting and climate modelling respectively. Concluding this thesis, the final chapter summarizes the main outcomes, sets them into context, and gives an outlook on possible future extensions.

## 1.1 Precipitation

Precipitation, in the form of rain, snow, or hail, results from the condensation of water vapour in the atmosphere due to a complex interplay between physical processes on a range of spatial and temporal scales. The multi-scale aspect and the many processes involved make numerical precipitation modelling and accurate observational measurements challenging. Accurate observations, however, form the basis for developing and validating new models, such as the post-processing methods in this thesis. The following sections are thus meant to give a brief introduction to these topics.

### 1.1.1 Physical processes

In essence, precipitation occurs when air rises and thereby cools due to the expansion in the lowered pressure, leading to condensation of the moisture and the formation of larger water droplets. The droplets then fall down under the gravitational pull. The generation of precipitation thus requires three main mechanisms (Trenberth et al., 2003):

**(i) The source for the uplift** Causes for the upward motion of the air can be manifold. For example, atmospheric instabilities range from small turbulent and convective motion to planetary Rossby waves (Mo and Higgins, 1998; Boers et al., 2019). The orography can also cause uplifts, e.g., when wind flows over mountain ranges (Houze Jr., 2012; Trenberth et al., 2003).

**(ii) The microphysical condensation process** The smallest processes involved in the generation of precipitation are related to the condensation in clouds. Water droplets grow by micro-physical interactions until they are heavy enough or undergo a phase shift to fall. Further, aerosols such as those emitted by humans can interact and affect the formation process (Rotstayn and Lohmann, 2002; Trenberth et al., 2003; Harrison et al., 2020).

**(iii) The availability of moisture** The moisture availability depends on the local thermodynamic conditions and the advective transport to or from other regions. The ratio of the two sources is called *recycling* ratio and depends on the season and region (Dai, 2001). Especially storms and cyclones that generate large amounts of precipitation draw moisture from the surrounding atmosphere leading to larger precipitation sums than possible from the locally available moisture (Trenberth et al., 2003).

To better understand the change in large-scale moisture available for precipitation with a warming of the atmosphere, the Clausius Clapeyron relation of the saturation water vapour pressure can be used (Berg et al., 2013; Lehmann et al., 2015; Guerreiro et al., 2018),

$$\frac{d \ln e_s}{dT} = \frac{L}{RT^2}, \quad (1.1)$$

where  $e_s$  is the saturation water vapour pressure<sup>1</sup>,  $L$  the latent heat of vaporization and  $R$  the ideal gas constant. For temperatures in the lower layers of the atmosphere, the right-hand side of Eq. 1.1 has a value of about 0.07, which corresponds to an increase in  $e_s$  of about 7% per 1 degree [K] warming (Held and Soden, 2006). This scaling has also been found to hold for rainfall on near-global averages in observations (Traxl et al., 2021).

On regional scales, however, the trends in precipitation can vary strongly (Ali et al., 2018; Traxl et al., 2021). Regions where precipitation is dominantly generated by convection have been found to exhibit trends that exceed the thermodynamic scaling expected from Eq. 1.1. Here, precipitation extremes are thus projected to increase in frequency and severity in the coming decades of anthropogenic warming (O’Gorman and Dwyer, 2018). On the other hand, some regions show negative precipitation trends, for example, due to the decreasing availability of moisture (Ali et al., 2018).

---

<sup>1</sup>Sometimes referred to as the water holding capacity of the atmosphere.

Therefore, reliable projections of long-term changes in precipitation for future warming scenarios are critical. Similarly, accurate weather forecasts of extreme events will likely gain further importance with a warming of the atmosphere.

### 1.1.2 Statistical characteristics

The main characteristics that set precipitation apart from many other atmospheric variables and make it challenging to predict and model are the strongly skewed non-Gaussian distribution and the discrete separation between dry and wet conditions, i.e., zero or finite precipitation sums (Koutsoyiannis, 2004a,b). Another defining characteristic is the substantial variability of precipitation in space and time, called intermittency. It describes the sparseness of the precipitation signal and can be quantified as the ratio between precipitation and no-precipitation events for a given precipitation threshold and time interval (Dunkerley, 2015).

As described in section 1.1.1, precipitation results from the interaction between a large range of physical processes on various temporal and spatial scales. Hence, depending on the spatial and temporal scale of interest, the dominating physical processes might change and, in turn, will lead to different statistical characteristics. For example, atmospheric convection can generate localized bursts of large precipitation amounts. In contrast, large-scale and more uniformly distributed precipitation typically occurs in frontal systems (Pfahl et al., 2017).

The intermittency of precipitation events has been found to increase with time scales from minutes to days (Dunkerley, 2015; Schleiss, 2018). On monthly to annual time scales, the intermittent signal is decreasing again due to the averaging effect (Schleiss, 2018), thus exhibiting a dependence on the time scale.

Several studies have investigated the scaling behaviour of precipitation time series across temporal and spatial scales using conceptual models and observations (Veneziano et al., 2006; Claussnitzer, 2010) with some indication of self-similar scaling for certain ranges of scales. This can be physically motivated by a similar scale invariance of turbulent eddies in the atmosphere interlinked with precipitation (Schertzer and Lovejoy, 1987).

A geometrical way to quantify how patterns change with the scale at which they are measured was introduced by Mandelbrot (1967) with the notion of the *fractal dimension*. The concept derives from how shapes or patterns change in different dimensions under scaling. Intuitively, the dimension of a pattern can be related to the number of squares  $N$  with side length  $s$  that are required to cover it (Lovejoy et al., 1987), as

$$N \propto s^{-D}. \quad (1.2)$$

For example, a single two-dimensional square ( $D = 2$ ) of side length  $s = 1$  can be divided into  $N = 4$  smaller squares of side length  $s = 0.5$ , or  $N = 16$  squares of side length  $s = 0.25$  and so on.



To find the fractal dimension of shapes that lie between integer dimensions, the box-counting method can be used (Meisel et al., 1992). It divides the pattern or shape to be analysed into measuring “boxes” of side length  $s$  and counts the number of boxes that cover the pattern  $N$ . From these measurements, the dimension  $D$  can then be estimated with Eq. 1.2, i.e.,

$$D = \frac{\log(N)}{\log(1/s)}. \quad (1.3)$$

The physical processes that generate precipitation change with temporal and spatial scales can also vary strongly geographically and seasonally (Ali et al., 2018; Traxl et al., 2021). This adds another layer of complexity when studying or forecasting precipitation.

Another key challenge is capturing the statistics of rare and extreme events in the far end of the upper tail in the distribution, given their severe impact. This is especially important for assessing and predicting the risk of such extreme events, e.g., for engineering tasks or disaster prevention and mitigation measures.

**Extremes** There are various definitions for extreme events, of which two are the most common ones in the context of extreme value theory (EVT). In the first, the maximum values of finite sequences or blocks, so-called *block maxima* (BM), are taken as extremes, e.g., the maximum precipitation sum in a day, month, or year (Serinaldi and Kilsby, 2014; Lehmann et al., 2015). In the second approach, called *peak over threshold* (POT), values in a continuous record exceeding a certain threshold, e.g., the 90th, 95th, or 99th percentile of the empirical distribution, are taken as extremes (Boers et al., 2019). BMs can be modelled with the generalized extreme value (GEV) distribution, while the distribution of POTs can be described by a generalized Pareto (GP) distribution (Serinaldi and Kilsby, 2014).

Accurately estimating the distribution of relative precipitation frequencies can be challenging in itself, however. Empirical histograms can be used, for example, to estimate the distribution of past observations, but do not allow extrapolation to future unseen events. On the other hand, parametric distributions require an assumption of the distribution form. Estimating the distribution parameters to fit the tails can be challenging due to the small sample size (Frei and Schär, 2001; Koutsoyiannis, 2004a). Hence, uncertainties about the occurrence probability of extremes remain large.

### 1.1.3 Observations

Accurate precipitation observations are essential in nowcasting applications (Franch et al., 2020a), in data assimilation for numerical weather predictions (Geer et al., 2018), or the validation of numerical weather forecast and climate models (Kucera et al., 2013; Michaelides et al., 2009).

Following Tapiador et al. (2012) and Michaelides et al. (2009), this section aims to give an overview of the different ways to measure precipitation on the

ground (e.g. by rain gauges, disdrometers, or radar), or satellite-based (e.g. radar-based or passive sensors), and combinations with physical simulations.

**Ground-based measurements** There are three main tools for measuring precipitation on the ground, rain gauges, disdrometers, and radar.

Rain gauges provide a direct way of measuring precipitation sums at a specific location. By collecting the droplets that fall onto it, gauges mechanically count small volumes of precipitation (Strangeways, 2010). Disdrometers extend the approach by recording further properties, such as the physical phase and drop size distribution. (Tokay and Short, 1996).

Since gauges and disdrometers only provide local information, constructing datasets with a more extensive coverage requires the interpolation of station networks. This can be problematic, as the characteristics of precipitation vary with the spatial scale, depending on the relevant physical processes that generate it (Tapiador et al., 2012).

Weather radars on the ground can estimate precipitation over kilometer-scale areas and distinguish between different hydrometeor phases. This constitutes a valuable complement to local rain gauges and disdrometers, which must be placed in a high density to provide accurate spatial information. Ground-based measurements are unevenly distributed over the continents with higher densities in more developed regions of the world which can be a source of bias that has to be considered during the validation of global precipitation estimation products (Beck et al., 2017).

**Satellite-based measurements** Globally consistent estimates of precipitation are based on passive sensors or radar aboard satellites that orbit the Earth. Passive sensors can be grouped into three categories for visible (VIS), infrared (IR), and microwave (MW) spectral ranges.

Infrared-based methods aim to estimate the total precipitation on the ground by measuring the top-of-the-cloud temperature. Cold cloud temperatures are linked to vertical updrafts, which can cause precipitation. Visible light can be used during day time to infer additional information for the IR estimates to correct biases (Tapiador et al., 2012).

By measuring microwave radiation interacting with water particles in clouds or precipitation droplets, passive and active microwave techniques provide a more direct estimation than IR methods. At relatively long wavelengths compared to visible and infrared ranges, microwave radiation is not subject to atmospheric scattering and can pass through clouds.

Passive sensors suffer from the problem that critical information from the vertical distribution of atmospheric conditions cannot be inferred (Michaelides et al., 2009). Hence, cloud-resolving physical models are used to fill some of the gaps, but considerable uncertainties remain. Active MW estimations that provide their own radiation source were premiered in the Tropical Rainfall Measurement Mission (TRMM) (Kummerow et al., 1998; Huffman et al., 2007) and are continued in the Global Precipitation Measurement (GPM) mission (Hou et al., 2014).

**Combined precipitation products** Combining the different methods offers the opportunity to alleviate the limitation of one approach by using the advantages of another. For example, methods combining gauges and radar have been used to calibrate satellite-based measurements, such as in TRMM. Passive MW techniques can accurately capture the precipitation characteristics, while the IR sensor provides high-resolution information about the spatial distribution (Tapiador et al., 2012). IR sensors can also be used to estimate the temporal dynamics of the atmosphere, which can then be incorporated into physical advection models to improve the MW estimates.

*Reanalysis* products, such as ERA5 (Hersbach et al., 2020), combine ground and satellite-based observations with data assimilation routines from numerical weather prediction to derive comprehensive and consistent estimates of the atmosphere for a wide range of variables. The weather forecast model thus creates the best guess of the atmospheric state that optimally matches observations. These reanalyses are typically more consistent with well-observed variables. Variables such as precipitation, on the other hand, can exhibit notable deviations between reanalysis products (Beck et al., 2017, 2019; Hassler and Lauer, 2021).

## 1.2 Modelling weather and climate

The Earth system is incredibly complex, comprising numerous physical processes interacting on a large range of spatial and temporal scales. Modelling it numerically might seem like a daunting task. Yet, our ability to simulate weather and climate has steadily improved over the past decades (Bauer et al., 2015). Given the finite computational resources available, compromises to which processes can be included and resolved for a given modelling task are required. For example, forecasting the weather on times scales of hours to weeks is a very different problem than projecting climate scenarios for the next centuries. Further, global weather and climate models typically have different resolutions in space and time.

Weather models integrate the primitive equations on time steps in the order of minutes, while climate models can have longer time steps between hours to days (Drüke et al., 2021). The exact time step length can vary with the model component depending on the characteristic time scale of the dynamics. Global weather models have spatial grid cell sizes of around 10–50 kilometers (Benjamin et al., 2018; European Centre for Medium-Range Weather Forecasts, 2012), while global climate models simulate on coarser grids of around 25–200 kilometer horizontal resolution (Haarsma et al., 2016; Schneider et al., 2017b). The grid spacing in the vertical direction typically varies with height and can range between around 20 layers in low-resolution climate models in up to 140 layers in high-resolution weather forecasts (Drüke et al., 2021; Dunne et al., 2020).

Weather forecasting is essentially an initial value problem, i.e., determining the forecast’s initial conditions as closely as possible is crucial for the

predictive skill. Thus, much of the recent advances in numerical weather prediction can be attributed to advances in data assimilation techniques and ensemble forecasts of perturbed initial conditions. The modelling chain for numerical weather predictions is outlined in Fig. 1.1A.

The first step in the chain is concerned with gathering observations from which initial conditions for the numerical model can be derived via data-assimilation methods. Observations from heterogeneous sources such as weather stations, aeroplanes, and satellites are combined and used in four-dimensional variational (4D-Var) data assimilation (Navon, 2009) to compute optimal initial conditions. The numerical model then propagates the initial state forward in time, producing the actual forecast. Systematic forecast errors, e.g., caused by the parameterization of unresolved processes or due to missing processes, are corrected in a subsequent post-processing step. The final forecast product can then be provided to end users.

Modelling the Earth system on climate time scales, on the other hand, can be considered as a boundary condition problem. As described by Hoskins (2013) and depicted in Fig. 1.1B, in addition to the atmosphere, more and more components of the Earth system become important with increasing time scales. Hence, capturing their interaction and external forcings, such as greenhouse gas emissions or solar radiation, becomes increasingly important. Due to the chaotic nature of the atmosphere that causes exponential growth in the initial forecast error, deterministic forecasts are limited up to around two weeks. Therefore, accurate initial conditions are not relevant for climate projections. Post-processing also plays an important role in climate modelling, especially in assessing the impacts of anthropogenic global warming. These are simulated using impact models that are typically developed and calibrated with observation-based data but use ESM simulations as input for future scenarios. Hence, systematic errors in the ESM need to be corrected before the simulation output can be provided as input for the impact models.

This section aims to highlight the key differences between weather and climate simulations and the post-processing approaches in used both domains.

### 1.2.1 Numerical forecasts and projections

**Weather forecasting** Accurate weather predictions are important to many aspects of our society, from disaster prevention and mitigation of extreme events to energy management, agriculture, and transportation. Advances in computational technology, observational methods, and our scientific understanding have led to a steady improvement of forecast skill over the past decades, gaining about a day lead time per decade in medium-range forecasting (between about three to ten days) (Bauer et al., 2015).

As discovered by Lorenz in 1963 (Lorenz, 1963), there are fundamental limits to the predictability of the atmosphere due to the exponential growth in model errors, e.g., in the initial conditions or model formulation, that lead to diverging forecast trajectories. Hence, much of the advances in weather forecasting can be attributed to the assimilation of observations, improve-

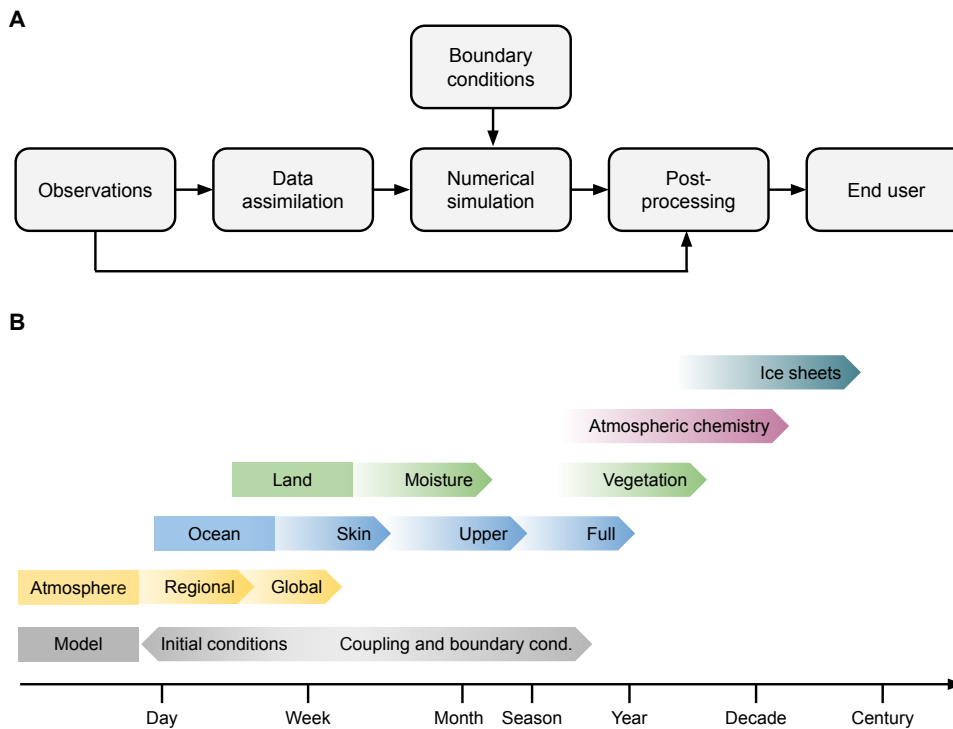


Figure 1.1: Components of the weather forecasting chain and time scales of Earth system model components. (A) The components of the weather forecasting chain, based on (Rasp, 2019). Observations are gathered from numerous sources, which are then processed to provide initial conditions for the numerical simulation in time. The final output of the numerical simulation is post-processed to remove systematic errors with respect to observations before being provided to an end user. (B) The different components of the ESM and their importance with respect to the time scales, adapted from (Hoskins, 2013). The atmosphere and oceans are the most important components for short to medium-range weather forecasting. At the same time, the provision of accurate initial conditions has a large impact on the forecast skill. On longer time scales, coupling and interaction of the model components become more important, including land, atmospheric chemistry, and ice sheets.

ments in the model formulations, and forecast uncertainty quantification using ensemble methods (Bauer et al., 2015).

An advantage of numerical weather prediction (NWP) as a scientific discipline is the ability to evaluate the models daily over the globe. Thus, improvements in the forecasting methods can be validated with high accuracy.

The spatial scale of events that can be skillfully forecasted is interlinked with the time scale of the forecasts (Hoskins, 2013). Short-term forecasts between hours and days are able to capture local small-scale events. Here, simulating atmospheric dynamics alone is often sufficient. High-impact weather events can be forecasted in the medium range of days to about two weeks. On this time scale, global weather simulations are required, and the coupling of the ocean with the atmosphere becomes important. Forecasts with lead times of months aim to predict large-scale weather patterns and regime changes. The interaction of land surface processes becomes important

at that time scale. Warming of the Pacific Ocean during the El Niño Southern Oscillation (ENSO) phenomenon is an example that can be predicted on seasonal time scales, typically below a year in advance (McPhaden et al., 2006; Ham et al., 2019).

With simulation time scales increasing towards decadal or centennial climate, the coupling of further components in the Earth system, such as the biosphere, including vegetation, atmospheric chemistry, and ice sheets, becomes more important. At the same time, accurate initial conditions cease to be relevant, as illustrated in Fig. 1.1B.

**Climate projections** Earth system models (ESMs) simulate the coupled components, such as the atmosphere, oceans, biosphere, and ice sheets, over periods of hundreds to thousands of years. Such models are built to answer fundamentally different questions than numerical weather prediction models. On long time scales, summary statistics and how they change due to the boundary conditions are the main question that ESMs are built to answer, rather than accurate forecasts of local weather. Examples are the change in global mean temperature with increasing CO<sub>2</sub> concentrations in the atmosphere or the irreversible tipping of ESM components into different states. The emphasis is, therefore, on the coupling and interaction of the different ESM components, while the initialization of the model mainly estimates its internal variability.

Due to the large uncertainty in the boundary conditions, such as future greenhouse gas emissions, different scenarios are often simulated and compared. In contrast to weather predictions, these experiments can often not be falsified, and the agreements of different model implementations across research institutions are used as an indication for their reliability<sup>2</sup> (Eden et al., 2012), for example, in the Coupled Model Intercomparison Project phase 6 (CMIP6).

However, today's most advanced and comprehensive ESMs exhibit large disagreement, e.g., in global mean temperature of about 1°C (Palmer and Stevens, 2019). Another prominent example is the equilibrium climate sensitivity (ECS), which measures the increase in global mean temperature due to a doubling of CO<sub>2</sub> in the atmosphere after reaching an equilibrium state (Schneider et al., 2017b; Balaji et al., 2022). Compared to observations over the past 50 years, ESMs exhibit significant biases. An example<sup>3</sup> is the overestimation of the precipitation band in the south of the Equator, the so-called double-peaked intertropical convergence zone (ITCZ) bias (Tian and Dong, 2020).

The main source for the large disagreement between different ESMs lies in

---

<sup>2</sup>However, as noted in (Eden et al., 2012), a high model agreement might not necessarily translate into high confidence, as simulations might agree due to common biases. Further, ESMs are rarely developed fully independently from each other.

<sup>3</sup>Further examples of ESM biases are the Amazon dryness, sea surface temperatures in the Southern Ocean, or differences in magnitude and frequency of the El Niño Southern Oscillation (ENSO) (Eyring et al., 2019; Hourdin et al., 2017).

their spatial and temporal resolution at which the physical differential equations are simulated. While ESMs have grown in complexity, including more and more processes, the horizontal spatial resolution is typically restricted to grid cells on the order of 25–200 kilometers. However, the microphysics of clouds and their feedbacks, which are crucial for ECS estimations and realistic precipitation distributions, require a minimum resolution of about 10 meters (Schneider et al., 2017b). Given the coarse resolution, turbulent convection of moisture, another important process for the local distribution of precipitation, cannot be resolved (Stevens and Bony, 2013). Such processes on sub-grid spatial scales are typically included as *parameterizations*, i.e., written as parameterized functions of the resolved variables.

**Sub-grid scale parameterizations** General circulations models (GCMs) that simulate the dynamics and energy transport of fluids in the atmosphere or oceans form the core of weather and climate models. The governing physical laws are formulated as coupled partial differential equations that are solved numerically on a discretized grid. The system of equations can be summarized as (Balaji et al., 2022),

$$\frac{\partial \mathbf{x}}{\partial t} = R(\mathbf{x}) + U(\mathbf{x}) + P(\mathbf{x}) + F, \quad (1.4)$$

where  $\mathbf{x}$  describes the system state, e.g., wind, water vapour, temperature, etc.,  $R(\mathbf{x})$  the resolved dynamics that can be written explicitly,  $U(\mathbf{x})$  the unresolved dynamics,  $P(\mathbf{x})$  the thermodynamic processes and  $F$  the external forcing. The “physics” of the system that acts on temporal or spatial scales which cannot be resolved with the given numerical discretization is represented by  $U(\mathbf{x})$  and  $P(\mathbf{x})$ . It thus has to be parameterized (Balaji et al., 2022), e.g., as

$$U(\mathbf{x}) + P(\mathbf{x}) = \sum_{\theta} \mathcal{M}(\mathbf{x}, \theta), \quad (1.5)$$

where  $\theta$  is a set of parameters derived from theoretical considerations or empirical studies. Increasing the resolution of the numerical model allows an extension of the number of processes that can be explicitly modelled and therefore reduces the need for parameterizations. This, however, comes at an increased computational cost, which limits the periods over which simulations can be performed or the size of model ensembles used for uncertainty quantification. Since relevant processes in the Earth system are acting on spatial scales as small as micrometers (Balaji et al., 2022), a feasible target resolution where all important processes are resolved will most likely be forever out of reach.

## 1.2.2 Statistical post-processing

Numerical models for weather prediction and climate projections often exhibit systematic biases. One important source is the subgrid-scale parameterization, as discussed in the previous section 1.2.1. By approximating the effect of

small-scale processes on large-scale variables, errors are introduced. Statistical post-processing methods have been developed to correct these by comparing the model simulations with observations.

In the following, an overview of different post-processing methods for both weather forecasts and climate projections is given.

**Post-processing weather forecasts** Statistical post-processing methods for weather forecasting have been developed for almost as long as numerical weather models themselves (Wilks and Vannitsem, 2018).

There are two main sources of uncertainties and errors in numerical weather forecasts. The first results from the dependence of the deterministic predictability on the initial conditions. Since the true initial conditions cannot be determined exactly in the real world, the choice of initial conditions may strongly affect the forecast performance. The sensitivity on the initial conditions is further state-dependent, i.e., varies over time, and therefore needs to be estimated for each forecast individually (Lorenz, 1963; Wilks and Vannitsem, 2018).

This has led to the development of ensemble forecasts that extend single deterministic forecasts to probabilistic predictions (Lorenz, 1965; Palmer, 1993). The uncertainty in the initial conditions is thereby modelled as a distribution over the ensemble with perturbed initial conditions. In other words, the distribution over the initial conditions is propagated in time by numerically integrating each weather model in the ensemble.

The resulting predictive ensemble distribution, however, can also be biased. This can be caused by an imperfect selection of the initial distributions and imperfect models. The second source of uncertainties is the incomplete model due to the finite resolution in space and time, together with the limited number of modelled processes. Post-processing in the weather prediction context can be applied to correct errors in single forecasts or to improve probabilistic ensemble predictions.

Probabilistic post-processing methods aim to improve properties such as the sharpness and calibration of the predictive distribution. Sharpness can be seen as the concentration of the distribution, while calibration refers to the agreement between forecast probability to the observed event frequency (Gneiting et al., 2007; Lerch et al., 2017).

Methods for correcting single weather forecasts are typically referred to as *model output statistics* (MOS)<sup>4</sup> (Glahn and Lowry, 1972). Especially for near-surface or surface variables, systematic biases in numerical weather forecasts can be significant (Wilks and Vannitsem, 2018). Due to the limited resolution, complex boundary effects, e.g., due to the orography, surface types, etc., cannot be resolved. Similarly, small-scale processes, such as microphysics in clouds, remain unresolved and must be approximated in parameterizations. One of the earliest MOS methods is based on a linear regression model that predicts a target variable  $\hat{y}(t)$  from given feature variables  $x_i(t)$ ,  $i = 1, \dots, N$

<sup>4</sup>However, sometimes ensemble post-processing methods are also referred to as MOS in the literature.



at time  $t$  that the weather model forecasts with

$$\hat{y}(t) = a_0 + a_1x_1(t) + a_2x_2(t) + \dots + a_Nx_N(t), \quad (1.6)$$

where  $a_i$  are the regression coefficients, e.g., determined by minimizing the mean squared error (MSE) between  $\hat{y}$  and the observation  $y(t)$  (Glahn and Lowry, 1972). This approach is suitable for weather forecasts, as the prediction  $\hat{y}(t)$  can be directly compared to a ground truth observation  $y(t)$ , e.g., with metrics such as the MSE. Publication P1 extends this approach using a non-linear regression model in the form of a convolutional neural network to improve predictions of heavy rainfall events.

**Post-processing climate projections** Earth system models for climate simulations have three main sources of uncertainty (Eden et al., 2012). The first source is the uncertainty in the external forcing, e.g.,  $F$  in Eq. 1.4, and how the model responds to it, e.g., to solar radiative forcing or CO<sub>2</sub> emission scenarios. The second source is related to the internal variability of the ESM model, which is caused by the interaction and feedback between processes. Parameterizations of small-scale processes that cannot be resolved, e.g.,  $U(\mathbf{x})$  and  $P(\mathbf{x})$  in Eq. 1.4, are the third source for uncertainty. This is particularly important for precipitation that is generated by parameterized microphysical processes.

To estimate these uncertainties in ESMs, ensemble runs are typically performed by perturbing the initial conditions or the parameterizations that lead to varying internal variability. The initial conditions for climate model runs are thereby not based on observations as in weather forecasts, but for example, based on control runs with preindustrial conditions. Still, the perturbation leads to different internal variability modes of the model that can be used for uncertainty quantification. The uncertainty of future climate scenarios is typically assessed by applying different external forcings.

Using historical observations, methods have been developed to reduce these uncertainties. Techniques such as emergent constraints aim to select or weight models in an ensemble that lie closer to observations (Hall et al., 2019; Williamson et al., 2020; Knutti et al., 2017). For the model error and uncertainty introduced by parameterizations, post-processing methods can be applied that reduce the model bias with respect to observations.

The correction of systematic bias with respect to observations through post-processing is especially important for impact modelling. Impact models assess the effect of anthropogenic global warming, for example, on flooding events, sea level rise in coastal regions, or crop yields and vegetation changes. These impact models are, however, developed and calibrated using observational input data. This way, the modelled and observed impacts can be directly compared and evaluated. For future impacts, projection simulations from comprehensive ESMs have to be used as input. This can cause severe inconsistencies as the ESM output fields often exhibit very different characteristics than the observations. Hence, post-processing methods aim to

bridge this gap and provide impact models with future scenario simulations that are more consistent with observations.

Due to the chaotic nature of the Earth system, the “weather trajectory” of the climate simulation does not follow the observed state closely, as errors in the initial conditions and model grow exponentially in the first few days of the simulation. This is not a model deficiency but rather a natural consequence of the chaotic system. Hence, ESMs are built to produce realistic summary statistics aggregated over long periods, e.g., decades or centuries. This also means that the post-processing aimed to reduce the model bias cannot follow the MOS approach in Eq. 1.6 since comparing variables locally for each time step  $t$  on “weather scales” is not meaningful.

Therefore, ESM bias correction methods are typically designed to compare and optimize *distributions* over time. One particularly successful method is *quantile mapping* (QM). It can be used to correct the entire distribution of a modelled variable  $x_{\text{model}}(t)$  at a given grid cell (Cannon et al., 2015). This is done by approximating a function  $f(x)$  that maps the value of  $x_{\text{model}}(t)$  to a corresponding value in the observations  $x_{\text{obs}}(t)$ , i.e.,

$$x_{\text{obs}}(t) \approx \hat{x}_{\text{model}}(t) = f(x_{\text{model}}(t)). \quad (1.7)$$

The function  $f(x)$  can be constructed from the cumulative distribution functions (CDFs), that are estimated over a historical period for both observational and modelled time series with

$$\hat{x}_{\text{model}}(t) = \text{CDF}_{\text{obs,hist}}^{-1} \{ \text{CDF}_{\text{model,hist}} [x_{\text{model}}(t)] \}, \quad (1.8)$$

where  $\text{CDF}_{\text{obs,hist}}^{-1}$  is the inverse of the CDF (i.e. quantile function) of the observations. Quantile mapping methods differ in their way of estimating the CDFs. Empirical quantile mapping uses histograms to approximate the CDFs without making assumptions about the shape of the distribution. On the other hand, parametric quantile mapping assumes a functional form of the distributions. It has the main advantage that values can be extrapolated if they fall outside the historical range.

Bias correction ESM climate simulations w.r.t. observations assumes that the model bias is stationary. In other words, the correction function derived from historical ESM simulations and observations, e.g.,  $f(x(t))$  in Eq. 1.7, is assumed to be invariant in time (Maraun, 2012). While the assumption might not be fully justified (Christensen et al., 2008), it is difficult to estimate the degree of non-stationarity in the model bias, given the small climate change signal in the period where sufficiently dense historical observations are available (Maraun, 2012). Generally, a trade-off has to be considered between reducing the ESM bias w.r.t. historical observations for better impact modelling and, on the other hand, the risk of underestimating the transient climate change signal.

Since the long-term climate of the Earth system is non-stationary, quantile mapping methods have been developed to preserve the transient trends. For example, this can be done in three steps, by first detrending the ESM

simulation, then applying the bias correction on the detrended data, e.g., with Eq. 1.8, and finally adding the trend to the corrected simulation again (Cannon et al., 2015; Lange, 2019).

The non-stationarity of the climate is also challenging for deep learning-based post-processing methods that typically assume identically distributed training data. The following section introduces key machine learning concepts, particularly for training deep neural networks from computer vision and image processing for post-processing tasks.

### 1.3 Deep learning for post-processing

Deep learning, the field of machine learning that is based on deep artificial neural networks (ANNs), has gone through different periods of popularity and development since the introduction of the multi-layer perceptron (MLP) model in the 1950s (Rosenblatt, 1959; Schmidhuber, 2015; Goodfellow et al., 2016).

The surge in popularity in recent years can be attributed to the increasing availability of training data, combined with advances in dedicated hardware such as GPUs and TPUs, as well as developments in algorithms and software that allow training ANNs and processing datasets of increasing size. Build with inspirations from neuroscience; ANN architectures have become more and more complex as have the tasks they can solve (LeCun et al., 2015; Goodfellow et al., 2016).

This section introduces basic concepts of machine learning and neural networks, particularly for computer vision and image processing applications that can use spatial correlations in gridded data such as images or, in the case of this thesis, geographical fields from numerical simulations. Further, metrics that can be used to quantify image characteristics, such as blurring, are discussed together with suitable learning approaches. The following is largely based on (Goodfellow et al., 2016).

#### 1.3.1 Machine learning basics

**Learning approaches** Machine learning algorithms can be roughly grouped into three categories of supervised, unsupervised, and reinforcement learning. The categories differentiate between the tasks the algorithm has to solve and the nature of the training data, i.e., whether the training data contains input features and labelled targets.

Supervised algorithms need labelled training data in the form of inputs (“features”) and outputs (also called “targets” or “labels”). For example, in image recognition, the features are the input images, and labels are assigned to the object(s) in the image, which the algorithm is tasked to predict. Another example is regression, where a continuous target variable must be predicted from a set of input features.

Unsupervised learning, roughly speaking, tries to identify structures in unlabeled data. Examples are clustering algorithms that group features based on their similarity or dimensionality reduction methods. Generative models are another example of unsupervised learning that try to approximate the distribution of the training data, s.t. new samples can then be drawn (Goodfellow et al., 2020).

However, combinations of the two categories are possible as well. Generative adversarial networks (Goodfellow et al., 2014) learn to generate samples that are indistinguishable from the target data distribution by including elements of supervised learning in the form of a second discriminator network that is trained as a classifier.

Reinforcement learning should be mentioned for completeness as another category of learning algorithms where an ML agent interacts with an environment. In this setting, the dataset is not fixed before training but accessed through interaction during training.

**Training techniques** After training an ML algorithm (in the following, also referred to as “model”), one usually measures its performance on new unseen data. Therefore, the dataset is divided into a training set used to optimize the model parameters (or “weights”) and a separate test set. The underlying assumption is that the samples in both sets are independent and identically distributed (i.i.d).

The assumption allows us to see the training samples as being produced by a single data-generating distribution and to quantify whether the model is *underfitting* or *overfitting*. Underfitting refers to the case where the model cannot reach a low error on the training set. On the other hand, overfitting occurs when the difference between the errors on the training and test set is large, i.e., when the training error is low but the test set error is high. This is illustrated in Fig. 1.2. To find the right balance between under- and overfitting, we can change the *capacity* of the model.

The capacity (or “expressiveness”) of a model determines which functions can be approximated, e.g., a regression model based on polynomials of degree two can learn the set of linear and quadratic functions. One way to increase the capacity is to increase the number of parameters. This could correspond to adding higher polynomial degrees in the given regression example. It allows the model to approximate a larger function space. However, it can make it harder to find the “true” data generating function from the large possible space, especially in the case of small training datasets, and can lead to overfitting. Decreasing the capacity too much, on the other hand, can also make it impossible to find the correct function, e.g., when approximating a non-linear function with a linear model, and lead to underfitting. Another technique to constrain the possible function space is *regularization*.

An extensive range of regularization techniques has been developed to control the models’ expressiveness. For example, a penalty in the form of an L1 or L2 norm on the weights can be added to the loss function that computes model error during training. Other approaches introduce sparsity

by removing parameters through pruning or dropout (Srivastava et al., 2014; Blalock et al., 2020). Simply increasing the size of the training data can also be seen as a form of regularization, as it adds further constraints to the model parameters. If additional training data can not be acquired, data augmentation methods, such as flipping or cropping images in computer vision problems, can be applied (Shorten and Khoshgoftaar, 2019). Adding additional terms to the loss function that capture different aspects of the tasks to be learned is another popular regularization strategy. Multi-task learning can be seen as such an approach, where the model has to learn several related tasks instead of a single (Caruana, 1997).

**Hyperparameters** The capacity of a model is determined through its *hyperparameters*. They control, e.g., the model architecture, the number of weights, or the type of regularization in a model. They also include parameters of the training method, such as learning rates (the “step size” of the parameter updates in gradient descent optimization) and batch sizes (the number of training samples used for an optimization step) or transformations in the pre-processing. A more detailed discussion of the model parameter optimization with gradient descent is given later in this section.

The choice of hyperparameters cannot be evaluated on the training set, as it would lead to a model with maximum capacity, and neither on the test set, as it is still part of the model optimization. Therefore, a third validation set is required for tuning hyperparameters.

The relation between under- and overfitting can also be seen from a statistical point of view as a bias-variance trade-off. From this perspective, the ML algorithm acts as a function estimator. Hence, the goal is to find an estimator  $\hat{f}(x)$ , e.g., with a neural network, of a target function  $y = f(x) + \epsilon$ , where  $\epsilon$  is some measurement noise, e.g., with  $\mathbb{E}(\epsilon) = 0$  and  $\text{Var}(\epsilon) = \sigma_\epsilon^2$ . Computing the mean squared error over the test set then gives insights into the trade-off between underfitting, leading to a low bias of the estimator, or overfitting resulting in a high variance of the estimator. This can be seen when decomposing the MSE as

$$\begin{aligned} \text{MSE} &= \mathbb{E}[(\hat{f}(x) - y)^2], \\ &= \mathbb{E}[\hat{f}^2(x)] + f^2(x) - 2\mathbb{E}[\hat{f}(x)]f(x) + \sigma_\epsilon^2, \\ &= \mathbb{E}[\hat{f}^2(x)] + f^2(x) - 2\mathbb{E}[\hat{f}(x)]f(x) + \sigma_\epsilon^2 + \underbrace{\mathbb{E}[\hat{f}(x)]^2 - \mathbb{E}[\hat{f}(x)]^2}_{=0}, \quad (1.9) \\ &= \mathbb{E}[\hat{f}(x)]^2 - 2\mathbb{E}[\hat{f}(x)]f(x) + f^2(x) + \mathbb{E}[\hat{f}^2(x)] - \mathbb{E}[\hat{f}(x)]^2 + \sigma_\epsilon^2, \\ &= \underbrace{(\mathbb{E}[\hat{f}(x)] - f(x))^2}_{(\text{Bias})^2} + \underbrace{\mathbb{E}[\hat{f}^2(x)] - \mathbb{E}[\hat{f}(x)]^2}_{\text{Variance}} + \sigma_\epsilon^2. \end{aligned}$$

The minimum MSE over a test set will thus have an optimal trade-off between bias and variance. This makes it a suitable loss function for many regression

problems. Observing the change in MSE over the epochs (iterations over the training data) during training allows determining an optimal stopping point of the process, i.e., early stopping of the training (Prechelt, 1998).

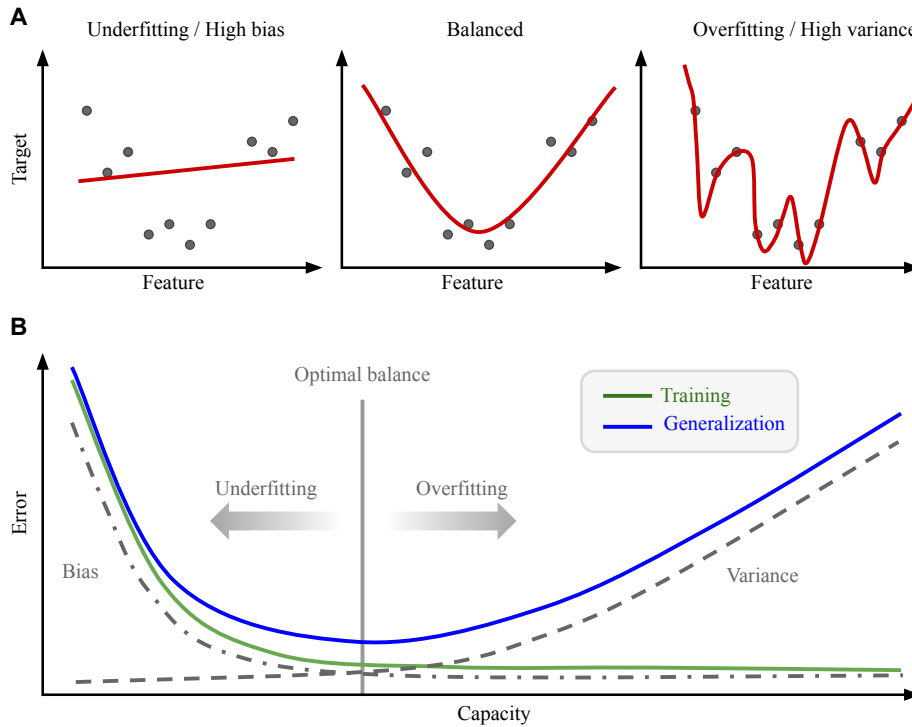


Figure 1.2: Sketch showing the trade-off between under- and overfitting, adapted from (Goodfellow et al., 2016). Increasing the capacity of a model reduces its bias and the risk of underfitting. After a certain point, the growing capacity can cause overfitting and high variance, leading to a high error on the test set. The goal is, therefore, to find the right balance between these two extreme cases.

**Optimization** The optimization procedure of neural networks is nowadays almost exclusively based on the *stochastic gradient descent* (SGD) algorithm. During training, the parameters  $\theta$  of a neural network  $f(x; \theta)$  are optimized with the goal of minimizing a cost function  $C(\theta)$ . The cost function measures the performance of the model and specifies the learning objective. The cost function is then constructed as the sum over a loss function  $\mathcal{L}$  evaluated on the training samples,

$$C(\theta) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(f(x_i; \theta), y_i), \quad (1.10)$$

where  $x_i$  and  $y_i$  are the  $i$ th feature and target sample, respectively, in a batch of  $m$  samples. Using batches of samples instead of the entire training set is computationally more efficient and typically used for training on large datasets in deep learning. The gradient of the cost function with respect to the model parameters is then estimated from a batch with,

$$\nabla_{\theta} C(\theta) = \frac{1}{m} \sum_{i=1}^m \nabla_{\theta} \mathcal{L}(f(\mathbf{x}_i; \theta), \mathbf{y}_i). \quad (1.11)$$

The new parameters  $\theta_{t+1}$  are then computed using gradient of the cost function w.r.t. the parameters,

$$\mathbf{g}_t = \nabla_{\theta} C(\theta_t), \quad (1.12)$$

with

$$\theta_{t+1} = \theta_t - \lambda \mathbf{g}_t, \quad (1.13)$$

where  $\lambda$  is the *learning rate*. In practice, the learning rate is often not fixed but gradually decreases during training with a scheduling algorithm. While the global minimum of the cost function is often not found, SGD typically can find a very low local minimum.

An extension of SDG optimization that is also used in the studies of this thesis is ADAM (Kingma and Ba, 2015). In ADAM, adaptive moments based on exponentially weighted averages of the gradient,  $\mathbf{m}_t$ , and the squared gradient,  $\mathbf{v}_t$ , are used to update the model parameters, i.e.,

$$\theta_t = \theta_{t-1} - \lambda \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t + \epsilon}}, \quad (1.14)$$

where  $\lambda$  is the learning rate again, and  $\epsilon = 1e^{-8}$  is a small constant for numerical stability.

**Back-propagation** In practice, the gradient of the cost function w.r.t the neural network’s weights in Eq. 1.11 is computed numerically with the *back-propagation* algorithm (Rumelhart et al., 1986).

It constructs a computational graph that stores variables as nodes and mathematical operations applied to these as edges. From the graph, the derivatives can be computed to get the total gradient using the chain rule of derivatives from calculus. The computational graph for back-propagation is found by storing the operations used to compute the output of the neural network from a given input in reverse order, i.e., starting from the network’s output and then going backwards layer by layer.

Computing the total gradient in the backward direction is computationally more efficient than computing it in the forward direction for each layer separately. The advantage is two-fold; it avoids duplicate computations because the gradient of a layer  $l$  does not depend on deeper layers  $l+1, l+2, \dots$  and does not need to cache intermediate results, since the full gradient w.r.t the cost function is computed for each layer directly. This is possible since the weights in a given network layer only change the loss by affecting the next layer and only using the information in the same layer.

Backpropagation is a special case of “reverse mode” automatic differentiation (AD) for a scalar-valued function such as the training cost. More generally, AD methods can also be applied to higher-dimensional functions.

### 1.3.2 Computer vision and image processing

The fields of computer vision and image processing are closely related, and a clear separation is not always possible. Broadly speaking, computer vision focuses on the abstract understanding of image or video data for tasks such as image classification, object recognition, segmentation, or video prediction, to name a few.

In image processing, the emphasis is more on analysing, transforming, and synthesising images for tasks such as denoising, compression, sharpening, or style transfer. Computer vision applications often include image processing techniques, e.g., for pre-processing, data augmentation, or evaluation in the workflow.

The development of convolutional neural networks and the ability to train them efficiently with back-propagation (LeCun et al., 2012) has revolutionized many parts of the two fields, where deep learning-based models attain state-of-the-art results.

**Convolutional neural networks** The convolutional neural network is inspired by the hierarchical structure of neuron layers in the visual cortex of humans and animals that process more complex patterns with increasing depth. It is one of the most successful architectures for gridded data such as images (LeCun et al., 2015). It is parameter efficient because individual neurons only process information in their limited receptive field. It further utilizes translation invariance, s.t., the position of the pattern in the input does not affect the response of the neurons.

At the heart of the CNN lies the convolution operation, which is typically implemented as a cross-correlation between an input  $x$  and a weighted kernel  $w$ , and in the case of a two-dimensional input image<sup>5</sup> has the form (Goodfellow et al., 2016)

$$y_{i,j} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x_{i+m,j+n} w_{m,n}, \quad (1.15)$$

where  $M$  and  $N$  are the height and width of the kernel in two dimensions and  $y_{i,j}$  is the value of the resulting feature map at position  $i$  and  $j$ . The resulting size, i.e., height or width, of  $y$  depends on three hyperparameters, the stride  $S$  that controls the step size between each convolutional operation, the padding  $P$  that extends the input image, and the kernel size  $K$ , leading to

$$\text{Size}(y) = \frac{I - K + 2P}{S} + 1, \quad (1.16)$$

where  $I$  is the size of the input image. Typically, network layers additionally apply a non-linear *activation function*<sup>6</sup>  $\sigma(\mathbf{y})$ , such as the rectified linear unit

<sup>5</sup>I.e., assuming an image with a single colour channel for simplicity.

<sup>6</sup>For a discussion on different activation functions and their properties, see, e.g., section 2.2 in Ray et al. (2023).



(ReLU), i.e.,  $\sigma(\mathbf{y}) = \max(0, \mathbf{y})$ , to produce the input of the next network layer  $\mathbf{z}$  with

$$\mathbf{z} = \sigma(\mathbf{y}). \quad (1.17)$$

CNNs are parameter-efficient architectures, typically having fewer parameters per layer than fully connected networks, with the size of the kernel being smaller than that of the input. This is motivated by the fact that patterns to be learned from images are typically smaller than the total image size. Moreover, parameters are shared because the same kernel parameters are used for the different locations in the input. Further, the extracted features will appear in the output feature map at a similar location.

Another concept that is central to CNNs is the *pooling* layer. Pooling is typically applied to aggregate information in a feature map, for example, by averaging over a pixel neighbourhood or computing its maximum value. It helps the network to learn translation-invariant patterns, i.e., where shifts in spatial directions of the input do not change the output significantly. Moreover, it allows the network to extract features at different spatial scales, as the pooling operation reduces the resolution, enabling kernels of the same size to extract larger spatial features in deeper network layers.

CNNs are typically constructed by stacking convolutional and pooling layers. Finding a suitable architecture is crucial for learning the desired function efficiently. For example, skip-connections (Ronneberger et al., 2015; He et al., 2016), which connect network layers by skipping some in between, can help the model to preserve high-frequencies in the image signal while also preventing problems such as vanishing gradients that can occur in deeper neural network architectures.

**Image processing and analysis** Besides the CNN architecture, the loss function primarily controls the function to be learned by the model.

A central question in this thesis is to find a suitable loss function that leads to desirable image properties in the CNN output. In the context of post-processing precipitation simulations of weather or climate models, resolving the characteristic high-frequency variability is particularly important.

The mean squared error (MSE),

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2, \quad (1.18)$$

between two variables  $x_i$  and  $y_i$ , is often used in supervised regression problems but has been found unreliable for detecting image distortions such as blurring (Wang et al., 2004). Therefore, extensions such as the structural similarity measure (SSIM) (Wang et al., 2004; Wang and Bovik, 2009; Rehman et al., 2013) have been proposed<sup>7</sup>. The SSIM compares two images  $\mathbf{x}$  and  $\mathbf{y}$

---

<sup>7</sup>See Fig. 2 in (Wang et al., 2004) for sensitivity comparison between MSE and SSIM for different image distortions such as blurring.

with respect to luminosity  $l(\mathbf{x}, \mathbf{y})$ , contrast  $c(\mathbf{x}, \mathbf{y})$  and structure  $s(\mathbf{x}, \mathbf{y})$ , as

$$\begin{aligned} l(\mathbf{x}, \mathbf{y}) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \\ c(\mathbf{x}, \mathbf{y}) &= \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \\ s(\mathbf{x}, \mathbf{y}) &= \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \end{aligned} \tag{1.19}$$

using the mean  $\mu$ , standard deviation  $\sigma$ , covariance  $\sigma_{xy}$  and small constants  $C_1$ ,  $C_2$ , and  $C_3$  for numerical stability. The SSIM then reads

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = [l(\mathbf{x}, \mathbf{y})]^\alpha [c(\mathbf{x}, \mathbf{y})]^\beta [s(\mathbf{x}, \mathbf{y})]^\gamma, \tag{1.20}$$

where  $\alpha$ ,  $\beta$ , and  $\gamma$  are parameters that can be adjusted to weight the different terms.

*Perceptual* image quality can also be quantified using CNNs. For example, by comparing the output or layer statistics of a trained image classification network, e.g., an Inception model (Szegedy et al., 2016) trained on the large ImageNet dataset (Deng et al., 2009), as done in the inception score (IS) and Fréchet inception distance (FID) (Salimans et al., 2016; Heusel et al., 2017). Generative adversarial networks (Goodfellow et al., 2014) similarly use a network-based metric in training, enabling it to produce realistic images of high-fidelity. Using a neural network instead of a heuristic or hand-engineered loss function that needs to be developed through many experiments makes these models attractive.

**Generative adversarial networks** As the name suggests, generative adversarial networks (GANs) are *generative* models, i.e., they can be trained to learn a target distribution from which samples can be drawn (or “generated”). They can generate highly realistic and sharp images (e.g. see Karras et al. (2019)). GANs do not require labelled training data and thus can be seen as a form of *unsupervised learning* (Goodfellow et al., 2020). A generative model aims to learn a distribution  $p_{\text{model}}(\mathbf{x})$  that approximates an unknown target distribution  $p_{\text{data}}(\mathbf{x})$  closely. Traditionally, this is done by fitting a parameterized distribution to the data drawn from  $p_{\text{data}}(\mathbf{x})$ , e.g., via maximum likelihood estimation. However, for high-dimensional distributions, this approach can become intractable (Goodfellow et al., 2020). GANs follow a different approach that only learns a transformation of samples from a given prior distribution  $p_z(z)$ , e.g., a Gaussian or uniform distribution, and therefore belongs to the class of *implicit* generative models.

The idea behind GANs is to learn the transformation of  $p_{\text{prior}}(z)$  through a “minimax” game between two deep neural networks (DNNs). One DNN takes the role of a so-called *generator*,  $G(z)$ , that learns the distribution  $p_{\text{model}}(\mathbf{x})$ . The other *discriminator* DNN,  $D(\mathbf{x})$ , learns the distinction between generated samples and such drawn from the target distribution  $p_{\text{data}}(\mathbf{x})$ , usually as a

classification task. The loss function for the GAN that was originally proposed by Goodfellow et al. (2014) is

$$\begin{aligned} \mathcal{L}(D, G) = & \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log(D(\mathbf{x}))] \\ & + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] \end{aligned} \quad (1.21)$$

The training then aims to solve

$$G^* = \arg \min_G \max_D \mathcal{L}(D, G), \quad (1.22)$$

to obtain the optimal generator network  $G^*$ . The discriminator is trained to maximize the loss in Eq. 1.21 by correctly classifying samples from the target distribution and generated ones. The generator tries to fool the discriminator by generating realistic samples that cannot be distinguished, thereby minimizing the loss.

In practice, the original GAN formulation in Eq. 1.21 is rarely used as it has been found to suffer from training instabilities and problems such as mode collapse, i.e., where the generator always produces the same sample (Arjovsky and Bottou, 2017). Thus alternatives for the loss function in Eq. 1.21 have been proposed (Salimans et al., 2016), such as the Wasserstein loss (Arjovsky et al., 2017), together with empirical techniques that can help to stabilize the training (Radford et al., 2016).

**Cycle-consistent GANs** The translation of unpaired images between two domains  $X$  and  $Y$  for style transfer tasks, e.g., translating photorealistic images to classical paintings or vice versa without changing their overall content, has been shown to be effective with cycle-consistent generative adversarial networks (CycleGANs) (Zhu et al., 2017). The CycleGAN uses two generator-discriminator pairs that learn inverse mappings between the two domains. The cycle-consistency constraint can then be defined as a regularizing loss term that computes the difference after one image translation cycle, i.e.,  $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$  and vice versa for  $y$ , using the L1 norm:

$$\begin{aligned} \mathcal{L}_{\text{cycle}}(G, F) = & \mathbb{E}_{x \sim p_x(x)} [||F(G(x)) - x||_1] \\ & + \mathbb{E}_{y \sim p_y(y)} [||G(F(y)) - y||_1], \end{aligned} \quad (1.23)$$

where  $G : X \rightarrow Y$  and  $F : Y \rightarrow X$  are the two generator networks. An identity constraint can additionally be imposed to regularize the networks to approximate an identity mapping with,

$$\begin{aligned} \mathcal{L}_{\text{ident}}(G, F) = & \mathbb{E}_{x \sim p_x(x)} [||F(x) - x||_1] \\ & + \mathbb{E}_{y \sim p_y(y)} [||G(y) - y||_1]. \end{aligned} \quad (1.24)$$

The log-likelihood loss in Eq. 1.21 is replaced by a mean squared error loss to improve the training stability, and the generator loss is reformulated in order to be minimized by the optimizer. The full generator loss then reads

$$\begin{aligned} \mathcal{L}_{\text{Generator}} = & \mathbb{E}_{x \sim p_x(x)} [(D_X(G(x)) - 1)^2] \\ & + \mathbb{E}_{y \sim p_y(y)} [(D_Y(F(y)) - 1)^2] \\ & + \lambda \mathcal{L}_{\text{cycle}}(G, F) + \tilde{\lambda} \mathcal{L}_{\text{ident}}(G, F), \end{aligned} \quad (1.25)$$

where  $\lambda$  and  $\tilde{\lambda}$  can be tuned to adjust the two loss terms. The discriminator networks are  $D_X$  and  $D_Y$  trained with

$$\mathcal{L}_{\text{Discriminator}} = \mathbb{E}_{y \sim p_y(y)}[(D_Y(y) - 1)^2] + \mathbb{E}_{x \sim p_x(x)}[(D_Y(G(x)))^2] \quad (1.26)$$

$$+ \mathbb{E}_{x \sim p_x(x)}[(D_X(x) - 1)^2] + \mathbb{E}_{y \sim p_y(y)}[(D_X(F(y)))^2]. \quad (1.27)$$

### 1.3.3 Combining domain knowledge with machine learning

Spatial precipitation fields from numerical weather and climate models often tend to be blurry, i.e., they cannot accurately resolve the variability in space. Therefore, a loss function based on a metric that penalizes blurriness in images can be advantageous for training a CNN in a post-processing setting.

Currently, research in atmospheric and Earth system science has started to apply deep learning methods for solving domain-specific problems due to their capability in text or image recognition, translation or generation (Reichstein et al., 2019; Huntingford et al., 2019; Schultz et al., 2021; Irrgang et al., 2021).

The following aims to give an overview of how data-driven methods from ML and DL can be combined with domain-specific knowledge for weather and climate modelling. In general, there exists a spectrum of different combinations with varying degrees of emphasis on either domain knowledge-based or data-driven approaches.

On the mostly knowledge based-side of the spectrum are physical models, typically in the form of (partial) differential equations that model the relevant physical processes. These models follow the physical laws of the system, e.g., conservation of energy or momentum. Unresolved processes can be included as approximations, typically in the form of parameterizations, where the free parameters can be calibrated with respect to observations. The number of free parameters is relatively small compared to large deep neural networks, which makes them relatively robust to generalize beyond the data used for calibration. The calibration procedure of the parameters is similar to ML optimization procedures (Cleary et al., 2021; Tsai et al., 2021). Process-based models are more interpretable than their deep learning-based counterparts, i.e., a physical meaning can usually be derived from the parameters. However, process-based models can usually only be applied to a specific problem for which they are designed.

On the opposite data-driven end of the spectrum, machine learning models, e.g., in the form of deep neural networks (DNNs), are trained to learn tasks such as emulating the dynamics of a system of interest purely from the training data. DNNs are rather flexible and can be applied to a range of problems. However, DNNs do not “know” about fundamental physical laws, such as the conservation of energy or mass, prior to training and might fail to learn them purely from data. The generalization to predictions outside the training data distribution can also be challenging. This is particularly important in the climate context, where historical observations are expected to have different distributions than future climate states. Further, with increasing

numbers of parameters, the models tend to become less interpretable (Molnar et al., 2020).

Between these two modelling paradigms lies a range of possibilities for combinations of the two. For example, purely DNN-based models can be tailored for more specific domain tasks, e.g., by extending their architecture to include physical equations and constraints, or differential equations-based models can be combined in so-called hybrid approaches to include data-driven components, e.g., in cases where the mathematical formulation of a process is not known or would be too expensive to compute.

**Incorporating domain knowledge into deep learning models** Using domain knowledge to make informed design choices of the deep learning model prior to training can have several advantages, especially when training data is limited. Selecting or designing a suitable architecture for the problem at hand helps to constrain the set of possible functions to be learned, e.g., by introducing inductive biases. Examples are the use of invariant properties that simplify the problem (Ling et al., 2016), developing architectures that fit the structure of the data, such as multiple scales that are common in atmospheric data or that are suitable for the spherical geometry of the Earth (Zhu et al., 2017; Weyn et al., 2020; Keisler, 2022; Pathak et al., 2022).

For problems where the data distributions for training and inference are identical, “soft” constraints in the form of additional regularization terms in the loss function can be imposed that penalize the violation of physical properties, e.g., in the context of “physics-informed” neural networks (Raissi et al., 2019; Greydanus et al., 2019; Li et al., 2022).

A central question in climate modelling is the generalization to out-of-sample predictions when training on historical observations. Here, architecture (or “hard”) constraints in the form of additional network layers can be included to ensure that conservation laws are fulfilled, which has been shown to improve the generalization (Beucler et al. (2021a)).

**Hybrid modeling** Hybrid modelling is a rather general term. Here, we will denote hybrid models as models that combine systems of physical differential equations with machine learning components. The ML component can be used to pre-process input data to the model (Kraft et al., 2022), emulate and replace physical components, e.g., subgrid-scale parameterizations in weather and climate models (Rasp et al., 2018; Gentine et al., 2018; Watt-Meyer et al., 2021; Yuval et al., 2021), or as post-processing to correct for biases in the physical model (Arcomano et al. (2022)).

An important distinction in hybrid modelling is the methodology for optimizing the hybrid model and particularly the ML component. One approach is so-called “offline” learning. Here, training data is usually generated by integrating the purely physical model in time. The resulting output data can then be used to train an ML algorithm. This approach is often practically convenient when the physical model cannot be accessed during training and therefore does not allow propagating gradients through the entire hybrid

model, which is necessary to update the network parameters. When including the offline trained ML algorithm in the physical model, a problem can be the out-of-sample predictions that result from the feedback of the ML component in the hybrid model and thus leads to different input data than seen during the offline training (Rasp et al., 2018; Sanford et al., 2022).

Therefore, recent efforts have begun to write physical models in modern programming languages such as Python or Julia (Schneider et al., 2017a; McGibbon et al., 2021; Kochkov et al., 2021; Häfner et al., 2021; Bauer et al., 2021; Kraft et al., 2022) that allow better integration of the ML component in the physical model.

**Interpretable and explainable AI** The performance of machine learning models typically increases with the model complexity and size (given sufficient training data and difficulty of the tasks to be learned) while the interpretability of the ML model decreases at the same time (Molnar et al., 2020). Deep neural networks are thus often considered as “black-box” models. Therefore, methods in the field of *interpretable* and *explainable* AI<sup>8</sup> are being developed to make them more transparent (Molnar et al., 2020).

There are several reasons why interpretability and explainability matter in ML applications to Earth system modelling. It can help to ensure that the model learns to identify the relevant physical processes. These insights, in turn, can be used to debug and improve the model, as well as increase the trust in its predictions (McGovern et al., 2019; Ebert-Uphoff and Hilburn, 2020). This is especially relevant for out-of-sample predictions in climate modelling, where predictions of the future often cannot be falsified (see section 1.2.1). Another application is ML-driven scientific discovery, e.g., where insights are extracted from an ML model trained on large volumes of data that might be infeasible to process by human experts (Karpatne et al., 2017; Zanna and Bolton, 2020).

There exists an abundance of different interpretability and explainability methods. Gradient-based techniques that leverage the differentiability of the model are useful for neural networks trained on image data. These techniques aim to attribute an “importance” score to input features with respect to the prediction. They can be grouped into methods that directly use the gradient of the model, such as saliency maps (Simonyan et al., 2014), or SmoothGrad (Smilkov et al., 2017) and methods that approximate the gradient, such as layer-wise relevant propagation (LRP) (Montavon et al., 2018), DeepLift (Shrikumar et al., 2019), or Integrated Gradients (Sundararajan et al., 2017).

The first group has the advantage that the gradient is always available, independent of the neural network architecture, whereas approximate gradients can be challenging to construct. Hence the latter is often used for simpler architectures, e.g., as in (Toms et al., 2020). On the other hand, using the

---

<sup>8</sup>The terms interpretability and explainability are often used interchangeably and are not clearly defined (Lipton, 2017). An attempt to do so and to distinguish these two terms can be found in Roscher et al. (2020).

gradient directly can be noisy as it can vary strongly on the local pixel level (Smilkov et al., 2017).

### 1.4 Outline of the thesis

To summarize the previous sections in preparation for the publications in the next chapters, a short outline of the thesis and its scope shall be given in the following.

The central topic that this thesis explores is the question of how deep learning methods from the computer vision and image processing domains can be applied to improve key characteristics of numerical precipitation simulations in a post-processing step.

Numerical models exhibit significant systematic errors in their simulation output that are largely due to the discretization of the model domain that requires approximations of small-scale processes as well as a limited model complexity. This is particularly important for precipitation since it results from the multi-scale interaction of numerous processes, including, e.g., small-scale microphysics in clouds.

Hence, numerical weather forecasts and climate projections exhibit various biases, e.g., by under- or overestimating the occurrence of rare extreme precipitation events or by producing overly smooth spatial fields that lack the defining small-scale variability, i.e., intermittency of precipitation patterns.

Classical statistical post-processing methods are typically applied to each location individually and are not designed to use spatial correlations efficiently. Therefore, these methods cannot improve spatial patterns effectively.

Convolutional neural networks, on the other hand, are built to use spatial correlation for tasks such as image-to-image translation, which is very similar to post-processing (or “translating”) spatial fields from numerical simulations using observations. Given these similarities, this thesis investigates suitable neural network architectures, loss functions and training techniques to improve weather forecasts and climate projections.

The following Chapter 2 explores such deep learning-based approaches in the weather prediction context. Here, the forecast skill of the numerical model or post-processing method can be directly evaluated by comparing the predicted field with observations for each grid cell and time instance. Supervised learning that requires such training samples can be applied using loss functions that improve the structural similarity of *paired* samples and can be modified to increase the predictive skill of rare extreme events.

The studies in Chapter 3 aim to improve spatial patterns and frequency distributions of numerical climate simulations, which poses different challenges. Small modelling errors grow exponentially in chaotic systems such as the atmosphere, which leads to deviating trajectories of the model simulation and observations. Therefore, training samples of daily precipitation fields become *unpaired*, which motivates the application of unsupervised learning techniques such as generative adversarial networks. Further challenging is

the non-stationarity of the Earth's climate, in which future projections become out-of-sample predictions for deep learning models trained on historical observations. Evaluating the similarity and realisticness of modelled spatial fields that are unpaired also requires different metrics than in short-term weather predictions.

The following main research questions are addressed in the following studies:

- Q1 *Can DNNs improve spatial patterns of precipitation, and in particular, the characteristic small-scale variability?* (Chap. 2 and 3)
- Q2 *Can DNNs accurately learn to correct the frequency distribution of precipitation extremes from biased model simulations?* (Chap. 2 and 3)
- Q3 *Can DNNs increase the forecast skill for rare precipitation events in numerical weather predictions?* (Chap. 2)
- Q4 *Can DNNs improve numerical climate simulations of precipitation?* (Chap. 3)



## 2 Post-processing rainfall forecasts with deep neural networks

*Based on*

### Deep Learning for Improving Numerical Weather Prediction of Heavy Rainfall

**Philipp Hess** and Niklas Boers, 2022.

Journal of Advances in Modeling Earth Systems, 14(3), e2021MS002765.

**Copyright** The article is published open access under the terms of the [Creative Commons Attribution-NonCommercial](#) license. The copyright remains with the authors.

**Contribution** PH and NB conceived the research and designed the study. PH performed the data processing, neural network training, and analysis. PH and NB interpreted and discussed the results. PH wrote the manuscript with input from NB.

**Summary** The study focuses on improving the numerical weather prediction (NWP) of rainfall and, in particular, of extreme events using a supervised deep learning approach.

The accurate prediction of rainfall and its extremes remains challenging due to the limited resolution and model complexity of the forecast model, which leaves important small-scale processes unresolved. This is particularly problematic for precipitation as it results from the interaction of processes across many scales, from continental weather systems down to cloud micro-physics.

In this study, a deep convolutional neural network based on the U-Net architecture ([Ronneberger et al., 2015](#)) is trained to reduce biases in the ensemble mean predictions of the Integrated Forecast System (IFS) ([European Centre for Medium-Range Weather Forecasts, 2019](#)) with respect to satellite-based observations from the Tropical Rainfall Measurement Mission (TRMM) ([Huffman et al., 2007](#)). The forecast and ground truth data have a high resolution of three hours in time and  $0.5^\circ$  horizontally in space with near-global coverage.

We show that a suitably designed loss function that combines a weighted mean-squared error to compensate for the strongly skewed frequency distribu-

tion with a multi-scale structural similarity index measure (MS-SSIM) (Wang et al., 2003) that is sensitive to blurring in images enables the network to accurately learn extreme events that are not provided by the NWP ensemble as input. Vertical wind velocities on different pressure levels are used as additional input features based on their physical link to heavy rainfall events.

The here-developed method can correct the relative frequency distribution with a comparable skill to state-of-the-art bias correction methods such as quantile mapping (Cannon et al., 2015) while achieving a better forecast skill. The latter is assessed using continuous metrics that cover the whole range of rainfall values and categorical skill scores that are commonly used in weather forecast evaluation for extreme events. The network trained with the novel loss function outperforms several other baseline methods, particularly in capturing the relative frequencies of rainfall sums in the distribution's upper tail.

### **2.1 P1 | Deep learning for improving numerical weather prediction of heavy rainfall**



Please turn to the next page.



## RESEARCH ARTICLE

10.1029/2021MS002765

# Deep Learning for Improving Numerical Weather Prediction of Heavy Rainfall

 Philipp Hess<sup>1,2</sup>  and Niklas Boers<sup>1,2,3</sup> 
<sup>1</sup>School of Engineering & Design, Technical University of Munich, Munich, Germany, <sup>2</sup>Potsdam Institute for Climate Impact Research (PIK), Potsdam, Germany, <sup>3</sup>Department of Mathematics and Global Systems Institute, University of Exeter, Exeter, UK

## Key Points:

- Correcting biases in the rainfall forecast of a numerical weather prediction ensemble with a deep neural network
- Training with a weighted loss function combining two terms enables the neural network to learn the heavy tailed target distribution
- The method improves the relative frequency and categorical skill scores of heavy rainfall

## Supporting Information:

Supporting Information may be found in the online version of this article.

## Correspondence to:

 P. Hess,  
[philipp.hess@tum.de](mailto:philipp.hess@tum.de)

## Citation:

 Hess, P., & Boers, N. (2022). Deep learning for improving numerical weather prediction of heavy rainfall. *Journal of Advances in Modeling Earth Systems*, 14, e2021MS002765. <https://doi.org/10.1029/2021MS002765>

Received 10 AUG 2021

Accepted 12 MAR 2022

**Abstract** The accurate prediction of rainfall, and in particular of the heaviest rainfall events, remains challenging for numerical weather prediction (NWP) models. This may be due to subgrid-scale parameterizations of processes that play a crucial role in the multi-scale dynamics generating rainfall, as well as the strongly intermittent nature and the highly skewed, non-Gaussian distribution of rainfall. Here we show that a U-Net-based deep neural network can learn heavy rainfall events from a NWP ensemble. A frequency-based weighting of the loss function is proposed to enable the learning of heavy rainfall events in the distributions' tails. We apply our framework in a post-processing step to correct for errors in the model-predicted rainfall. Our method yields a much more accurate representation of relative rainfall frequencies and improves the forecast skill of heavy rainfall events by factors ranging from two to above six, depending on the event magnitude.

**Plain Language Summary** Modeling rainfall is challenging because of its large variability in space and time, and its highly skewed distribution. Numerical weather prediction (NWP) models have to be simulated on discretized grids with finite resolution. Although important especially for the generation of rainfall, small-scale processes can therefore not be resolved explicitly and must be parameterized, that is, included as empirical functions of the resolved variables. This introduces model biases that can lead to an under- or overestimation of heavy rainfall events. Here we apply a deep neural network (DNN) to correct biases in the rainfall forecast of a NWP ensemble. The DNN is optimized with a loss function that includes weights to account for heavy rainfall events, and shows substantially improved performance in their prediction.

## 1. Introduction

Modeling and predicting rainfall, and in particular heavy rainfall events, remains is challenging. The relevant multi-scale dynamics range from small-scale droplet interactions to large-scale weather systems. Further, the high intermittency in space and time, as well the strongly non-Gaussian, right-skewed distribution (Koutsoyianis, 2004a, 2004b) make accurate predictions difficult.

The thermodynamic Clausius-Clapeyron relation (Allan & Soden, 2008; Donat et al., 2013; Guerreiro et al., 2018), and comprehensive model simulations (Masson-Delmotte, V. et al., 2021) suggest that the frequency and severity of heavy rainfall are expected to increase in a warming atmosphere (Fischer & Knutti, 2016). It should be noted, however, that the spatial patterns of these increases are expected to be heterogeneous and complex (Ali et al., 2018; Traxl et al., 2021). Correspondingly, accurate forecasts of heavy rainfall events will become ever more crucial for disaster prevention and mitigation.

Numerical weather prediction (NWP) models solve the fluid dynamical equations governing the dynamics of the atmosphere. They are essential for weather forecasting, including the prediction of heavy rainfall events. Despite the large improvements made over the past decades (Bauer et al., 2015), considerable sources of error remain in most of the models, in particular for rainfall (Boyle & Klein, 2010). Global NWP models, with a resolution of about 20 km, cannot explicitly resolve many of the relevant small-scale processes. These processes need to be included as sub-grid parameterizations, that is, they are written as functions of the explicitly resolved (grid-scale) variables. These parameterizations of important processes involved in the generation of rainfall introduces biases and errors that can lead to an under- or overestimation of the magnitudes of heavy rainfall events (Wilcox & Donner, 2007).

© 2022 The Authors. Journal of Advances in Modeling Earth Systems published by Wiley Periodicals LLC on behalf of American Geophysical Union. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Post-processing methods are commonly applied to the simulated model output to correct for such biases (Berg et al., 2012; Maraun, 2016; Wilks, 2011; Xu, 1999). Traditional approaches relate the biases to differences in long-term statistics of the simulated and observed variable. Among them, quantile mapping (QM) has become particularly popular for weather and climate model applications (Cannon et al., 2015; Déqué, 2007; Gudmundsson et al., 2012; Tong et al., 2021), as it allows to correct for biases over the entire distribution. While correcting the general long-term statistics, these methods, however, do not directly correct for spatial biases in synchronous events that are both modeled and observed.

Recent work has shown promising results by including data-driven machine learning methods including neural networks (LeCun et al., 2015), into the traditional NWP workflow. Well-suited applications of neural networks range from data-assimilation (Bocquet et al., 2020), purely data-driven and hybrid weather prediction and climate modeling (Brenowitz & Bretherton, 2019; Rasp et al., 2018; Rasp & Thuerey, 2021; Watt-Meyer et al., 2021; Weyn et al., 2020; Yuval & O’Gorman, 2020) to post-processing NWP output (Grönquist et al., 2021; Rasp & Lerch, 2018).

Here we correct the European Center for Medium-Range Weather Forecasting (ECMWF) (European Centre for Medium-Range Weather Forecasts, 2012) Integrated Forecast System (IFS) for biases in both general statistics and local events, by post-processing its rainfall output with a deep neural network (DNN).

When DNNs are tasked to infer a variable with large intermittency and a heavy-tailed distribution, such as rainfall, the optimization with the widely employed mean squared error (MSE) loss function often leads to a good approximate of the distribution's mean. By simply averaging over a sample batch, the loss is dominated by the most frequent values, while outliers in the tail of the target distribution only have a comparably small contribution. This can lead to blurring of the spatial patterns and a less accurate prediction of the high values in the tail, as the model focuses mainly on accurate predicting the most frequent values near the mean.

For rainfall, this problem has been addressed in different ways, for example, by translating the regression task into a classification problem (Agrawal et al., 2019; Sørnderby et al., 2020), by using methods from image quality assessment in computer vision (Tran & Song, 2019), and by employing a weighted loss function (Franch et al., 2020; Shi et al., 2017). The latter being composed of a weighted MSE and mean absolute error (MAE), with a set of five discrete weights determined by binned rainfall intensities. We show that the U-Net DNN architecture is able to infer high values in the far right tail of the target distribution from remotely sensed rainfall data. Notably, we use NWP ensemble simulations as input features, which do not exhibit an accurate representation of heavy rainfall events. To capture the heavy rainfall events and the intermittent spatial patterns, we introduce a new loss function, which combines a continuously weighted MSE with a structural similarity measure.

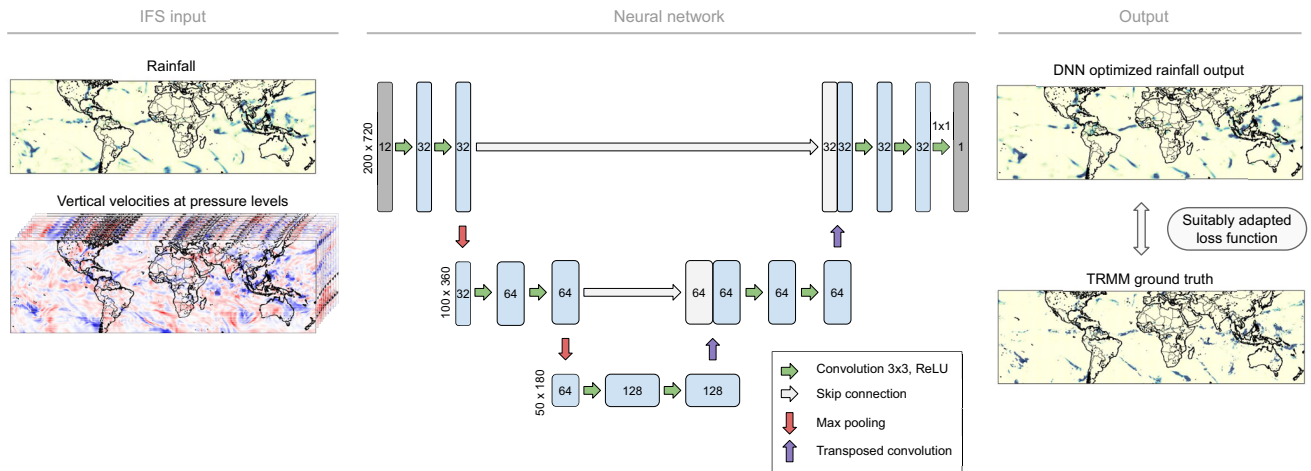
## 2. Materials and Methods

### 2.1. Integrated Forecast System

Atmospheric variables simulated as reforecasts by a ten-member ensemble of the IFS of the model cycle CY41R2 from the ECMWF (European Centre for Medium-Range Weather Forecasts, 2012) are taken as inputs of the DNN. The data is provided by the ECMWF at three-hourly time steps and  $0.5^\circ$  horizontal resolution. It is initialized twice daily at 06 and 18 UTC with a 12 hr lead time and small perturbations in the initial conditions. In this work, the ensemble mean of the variables is used, since taking the individual ensemble members as inputs would not be computationally feasible at present.

### 2.2. Training Data

The input features of the DNN are the three-hourly accumulated rainfall and vertical velocities of the IFS ensemble mean at the respective lead time. The forecast consists of three-hourly steps up to 12 hr lead time. The ensemble mean is taken from eleven pressure levels: 200, 250, 300, 400, 500, 600, 700, 800, 900, 950, and 1,000 hPa. The vertical velocity is dynamically linked to rainfall through convective processes and large-scale updrafts of warm, moist air (Müller et al., 2020; O’Gorman & Schneider, 2009; Pfahl et al., 2017). The satellite-based Tropical Rainfall Measurement Mission (TRMM) 3B42 V7 product (Huffman et al., 2007) is used as a training ground truth at three-hourly temporal resolution. Following (Beck et al., 2019; Rasp et al., 2020) the spatial resolution is regridded to  $0.5^\circ$  using bilinear interpolation to match the IFS grid. The TRMM data is considered to have high



**Figure 1.** Sketch of the U-Net-based deep neural network (DNN) architecture. IFS output for rainfall and vertical velocities is passed to the DNN, which produces rainfall output optimized to approximate corresponding spatial fields from a satellite-based, quasi-global high-resolution rainfall data set. The number of channels in the DNN is indicated inside each layer. The horizontal dimensions per pooling level are given on the left. The arrows show the operations applied after each layer. Green arrows indicate convolutional operations followed by a ReLU activation function. The skip-connections are shown as gray arrows, transferring the hidden state across the bottleneck. Orange and purple arrows indicate max pooling and transposed convolutions respectively. For a more detailed explanation of a similar sketch we refer the original U-Net publication (Ronneberger et al., 2015).

accuracy especially for heavy rainfall events (Boers et al., 2015). The geographic region of this study is the entire spatial coverage of the TRMM product, which ranges from 50° S to 50° N and 180° W to 180° W. Further, the June, July and August season is used and split into a training set of 8,096 samples (1998–2008), a validation set containing 2208 samples (2009–2011) to optimize the hyperparameters of the DNN model, and a test set with an equal number of samples for evaluation (2012–2014). Although the TRMM product is continued till present, a change of the satellites in 2014 has introduced significant biases, as shown in Figure S6 in Supporting Information S1, and the period after 2014 was therefore excluded.

### 2.3. Definition of Heavy Rainfall Events

We define heavy rainfall events as those 3-hourly time steps for which the rainfall sums exceed a pre-defined threshold. This threshold is determined individually for each grid cell in terms of percentiles. The percentiles are computed from the entire TRMM time series from 1998 to 2014 of 3-hourly time steps with rainfall amounts above 0.1 [mm/3h]. This allows to determine the event thresholds in the most accurate way by leveraging all the available data, which is important for the heavy rainfall events considered in this study.

### 2.4. Neural Network Architecture

The DNN architecture is based on the U-Net (Ronneberger et al., 2015), a convolutional neural network that can capture multi-scale spatial patterns. The U-Net includes a combination of pooling operations for large-scale feature extraction and skip-connections to preserve small-scale, high-frequency information. The U-Net architecture has shown good performance in weather prediction and post-processing tasks (Grönquist et al., 2021; Weyn et al., 2020). The model, shown in Figure 1, takes the standardized spatial fields of the atmospheric variables as input. The number of 12 input channels equals the number of variables times the corresponding number of pressure levels. The output layer has a single channel and spatial dimensions identical to the global rainfall grid. It applies a rectified linear unit (ReLU) to ensure non-negative output values. The number of weights per layer is reduced by half compared to the original model from (Ronneberger et al., 2015), and only two max pooling operations are found to be optimal for all the models in this study. This effectively reduces the model parameters size compared to the original U-Net. Adding more layers did not lead to improvements, as similarly found in (Grönquist et al., 2021; Weyn et al., 2020). The ADAM optimizer (Kingma & Ba, 2017) was employed for training the networks. We use a batch size of 64, an initial learning rate of  $10^{-4}$ , and early stopping with a patience of 20 epochs without improvement of the loss function on the validation data set to prevent overfitting. The learning

rate is reduced during training using a scheduler. It decreases by a factor of 0.1 after a period of 10 epochs without improvement on the validation loss.

## 2.5. Loss Function

To improve the training regarding high values and intermittency, we propose the weighted loss function

$$L_{\lambda}(y, \hat{y}) = \frac{\lambda}{N} \sum_{i=1}^N w(y_i) (y_i - \hat{y}_i)^2 + (1 - \lambda) \text{MS-SSIM}(y, \hat{y}), \quad (1)$$

where  $N$  is the number of training examples,  $w$  is a weight function and  $y$  and  $\hat{y}$  are the target and prediction, respectively. The cost function is thus a convex sum of the weighted MSE and the so-called multi-scale structural similarity measure MS-SSIM (Wang et al., 2003) (named WMSE-MS-SSIM in the following), introducing an additional hyperparameter  $\lambda$ . The weights  $w$  are defined as

$$w(y_i) = \min(\alpha e^{\beta y_i}, 1), \quad (2)$$

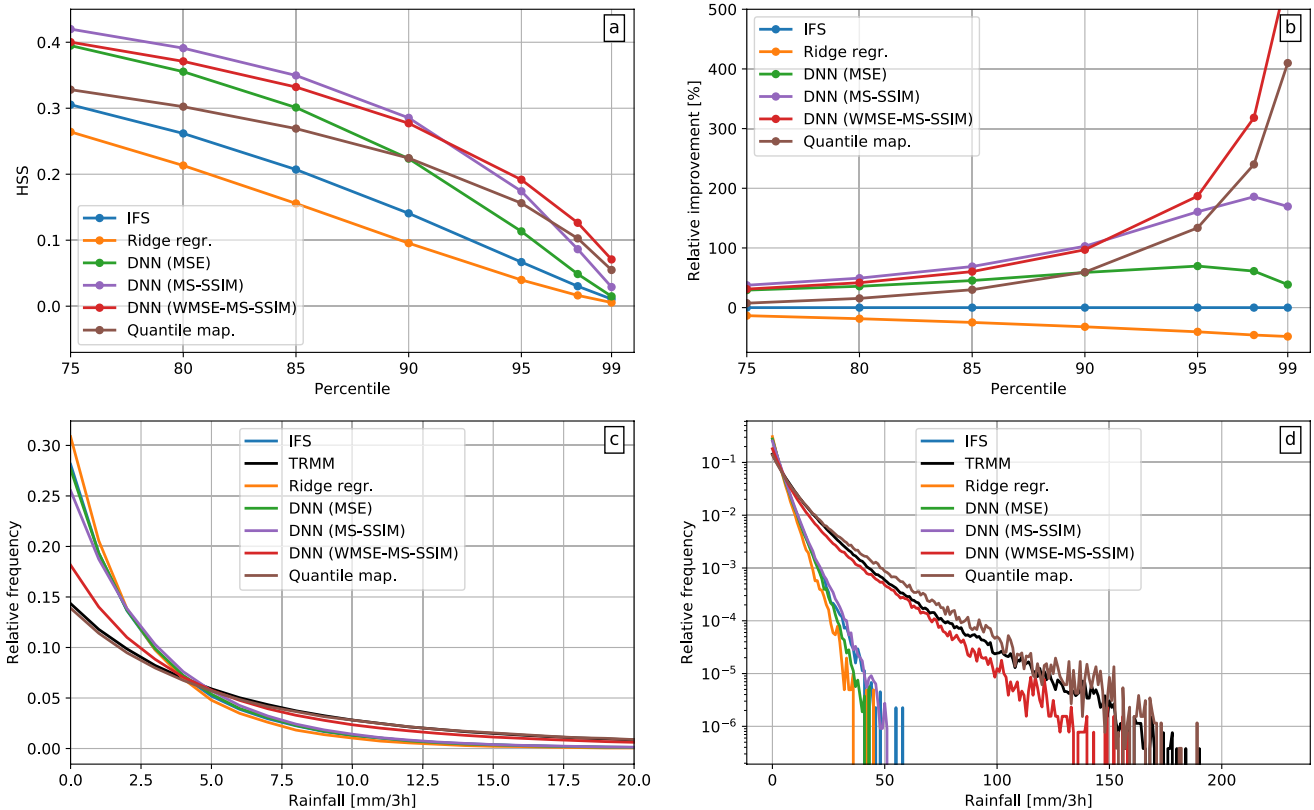
where  $\alpha$  and  $\beta$  are hyperparameters. We optimize all the network hyperparameters on the validation set using random search with uniform distributions for each loss function. The intervals of the parameter distribution were adapted during the optimization procedure. For the loss in Equation 1, we find through manual evaluation  $\alpha = 0.007$ ,  $\beta = 0.048$  and  $\lambda = 0.158$  to be optimal with respect to continuous metrics such as root mean square error (RMSE) and mean error (ME) as well as categorical skill scores such as F1 and CSI. Since the relative frequency of 3-hourly rainfall events decreases approximately exponentially with increasing magnitude, the weights aim to account for the statistical imbalance. Ebert-Uphoff et al. (Ebert-Uphoff & Hilburn, 2020) also use an exponentially weighted MSE loss to emphasise less frequent and high values when training a DNN to estimate radar composite reflectivity from satellite imagery. While the weighted MSE accounts for the skewed rainfall frequency distribution, the MS-SSIM evaluates the mean, standard deviation and covariance in the predicted rainfall output and ground truth. This is done through an iterative comparison of luminance, contrast and structure on different scales by downsampling and low-pass filtering the image signals (see supporting information). It is highly sensitive to blurring in images, as opposed to the MSE loss term. This can be seen for example, in Figure 2 in (Wang et al., 2004), showing a comparison of the sensitivity of MSE and MS-SSIM for different image distortions. Intuitively, one might hope that including the MS-SSIM will improve the spatial patterns of the DNN output, which is important for an accurate reproduction of heavy rainfall events. In our case, we indeed find that only optimizing with the weighted MSE leads to large biases, which can be removed through the addition of the MS-SSIM into the loss, with the role to improve the structural similarity. Further introducing bounds on the weights was crucial for a robust optimization of the network.

## 2.6. Baseline

We compare our method to two different baselines. A linear ridge regression (Hoerl & Kennard, 1970) with the IFS ensemble mean rainfall of a single grid-cell as input is used as the first baseline model. The regularization constant of  $10^{-3}$  was found to be optimal using the same validation method as for the DNNs. Including the vertical velocity fields did not improve the performance of this baseline model. In addition, we use QM (Déqué, 2007) as a second baseline. The period from 1998 to 2011 is used to estimate the cumulative distribution functions (CDFs) of the simulated  $F_{hist}$  and observed  $F_{obs}$  data with 750 discretized quantiles, which are found to be optimal. The CDFs are then used to match the corresponding quantiles via

$$\tilde{p}_{sim} = F_{obs}^{-1}(F_{hist}(p_{sim})). \quad (3)$$

Here,  $\tilde{p}_{sim}$  and  $p_{sim}$  are the quantile-mapping corrected and simulated rainfall values, respectively.



**Figure 2.** Relative rainfall frequencies and categorical heavy rainfall event forecast scores for the different post-processing models compared to the IFS. (a) The Heidke Skill Score (HSS) for events above increasing percentile thresholds is shown for the IFS (blue), ridge regression (orange), DNN trained with the mean squared error loss (green), the MS-SSIM loss (purple), with the WMSE-MS-SSIM loss proposed here (red) and quantile mapping (brown). A HSS greater than zero implies an improvement over a random forecast, and HSS = 1 would imply a perfect forecast (see supporting information). (b) The relative improvement of the HSS for the different machine learning methods over the IFS mean, is shown in percentages. Histograms of three-hourly rainfall event magnitudes are shown on a linear y-axis (c) and a logarithmic y-axis (d) for Tropical Rainfall Measurement Mission (black), IFS (blue), ridge regression (orange), DNN trained with the MSE loss (green), the MS-SSIM loss (purple) and the WMSE-MS-SSIM loss (red). The bins were chosen to be evenly spaced with a width of 1 mm/day.

### 3. Results

#### 3.1. Evaluation of the Continuous Forecast Skill of the Deep Learning Model

The evaluation results reported in the following are computed on the test data set. We first compare the histograms of the relative frequencies of the 3-hourly rainfall values for the outputs from IFS, the different post-processing models, and the ground truth given by the TRMM remote sensing product (Figures 2a and 2b). The histograms of grid-cell values are computed over the entire part of the globe covered by the TRMM data (50°S to 50°N) and test set period. Training the DNN with an MSE or a MS-SSIM loss leads to a similar rainfall frequency distribution as the IFS ensemble mean and the linear ridge regression baseline, with over-representation of low rainfall frequencies and underestimation of the tail, as compared to the observational TRMM target. Training with the WMSE-MS-SSIM loss function in Equation 1, instead, enables the DNN to infer a distribution that is substantially closer to the target distribution. The frequencies of low rainfall rates are correctly reduced, while at the same time achieving a better statistical representation of the heavy rainfall events in the tail. The ridge regression shows the largest bias toward low rainfall rates, hence not improving the IFS output at all. Applying QM to the IFS output on the other hand leads to an accurate representation of rainfall frequencies over the entire range of values - also for low values, as expected by construction.

We assess the continuous forecast skill of the different models by computing the RMSE, ME and the complex-wavelet structural similarity index (CW-SSIM; Sampat et al., 2009; see supporting information). The CW-SSIM allows a structural comparison of two images that is insensitive to small non-structural transformations such as rotation and translation, but sensitive to structural changes such as sharpness. Time steps with rainfall below

**Table 1**

*Continuous Validation Statistics Are Given for the Integrated Forecast System Ensemble Mean, Quantile Mapping, Ridge Regression, and the DNNs Trained With Different Loss Functions and the Input Variables Rainfall (P) and Vertical Velocity (W) From the IFS*

Model	Loss	Input	RMSE	%	ME	%	CW-SSIM	%
IFS	-	-	1.457	-	0.175	-	0.359	-
Quantile map.	-	P	2.071	-42.1	0.149	14.9	0.511	42.3
Ridge Regr.	MSE	P	1.473	-1.1	0.209	-19.4	0.359	0
DNN	MSE	W	1.375	5.6	0.165	5.7	0.388	8.1
DNN	MSE	P, W	1.372	5.8	0.166	5.1	0.395	10
DNN	MS-SSIM	P, W	<b>1.368</b>	<b>6.1</b>	0.136	22.3	0.441	22.8
DNN	WMSE-MS-SSIM	P, W	1.439	1.2	<b>0.135</b>	<b>22.9</b>	<b>0.545</b>	<b>51.8</b>

a threshold of 0.1 [mm/3h] have been excluded before applying the error metrics. Rainfall on such low scales cannot be measured accurately by satellite-based remote sensing (Huffman et al., 2007). Hence, completely dry times are not represented in the error statistics. The results are summarized in Table 1 as averages of the absolute cell-wise metrics. Training the DNN with the MS-SSIM leads to the lowest RMSE, while the WMSE-MS-SSIM loss function shows a ME similar to the MS-SSIM, and the highest structural similarity. Processing the IFS output with the ridge regression does not lead to improvements. Omitting rainfall from the input features and thus purely focusing on the vertical wind velocities W is not substantially affecting the performance of the model. The WMSE-MS-SSIM loss function combined with the MS-SSIM leads to an improvement of the ME by almost 23% and an improvement of the CW-SSIM metric by more than 50%. Besides the metrics discussed above, rainfall maps produced by the IFS, DNN and TRMM are shown in Figure S1 in Supporting Information S1 for a qualitative comparison. While QM is not able to reduce the RMSE of the IFS, it strongly reduces the ME and leads to high similarity values, reflected in the CW-SSIM.

### 3.2. Evaluation of the Forecast Skill of the Deep Learning Model for Heavy Rainfall Events

To evaluate the forecast skill for heavy rainfall events, categorical statistics can be computed from the contingency table containing the true positives and negatives, as well as the false positives and negatives (Table S1 in Supporting Information S1). A detailed definition of the events is given in Section 2.3 and the skill scores are defined in the Supporting Information. Table 2 summarizes the skill scores for events above the 95th percentile. The HSS, defined in the SI (Text S2 in Supporting Information S1), which is equal to zero for a random forecast and equal to one for a perfect forecast, is shown in Figure 2c for thresholds ranging from the 75th to the 99th percentile. Corresponding results for the other scores are given in the Figures S2 to S5 in Supporting Information S1. The DNNs improve the scores compared to the IFS mean and ridge regression, in particular for events above the 90th and higher percentiles (Figure 2c). Quantile mapping results in HSS, F1 and CSI values higher than for the MSE-trained DNN, but stays below the other two networks. While QM leads to a high probability of detection, it also shows a large FAR score indicating a high number of false positives. The DNN trained using the

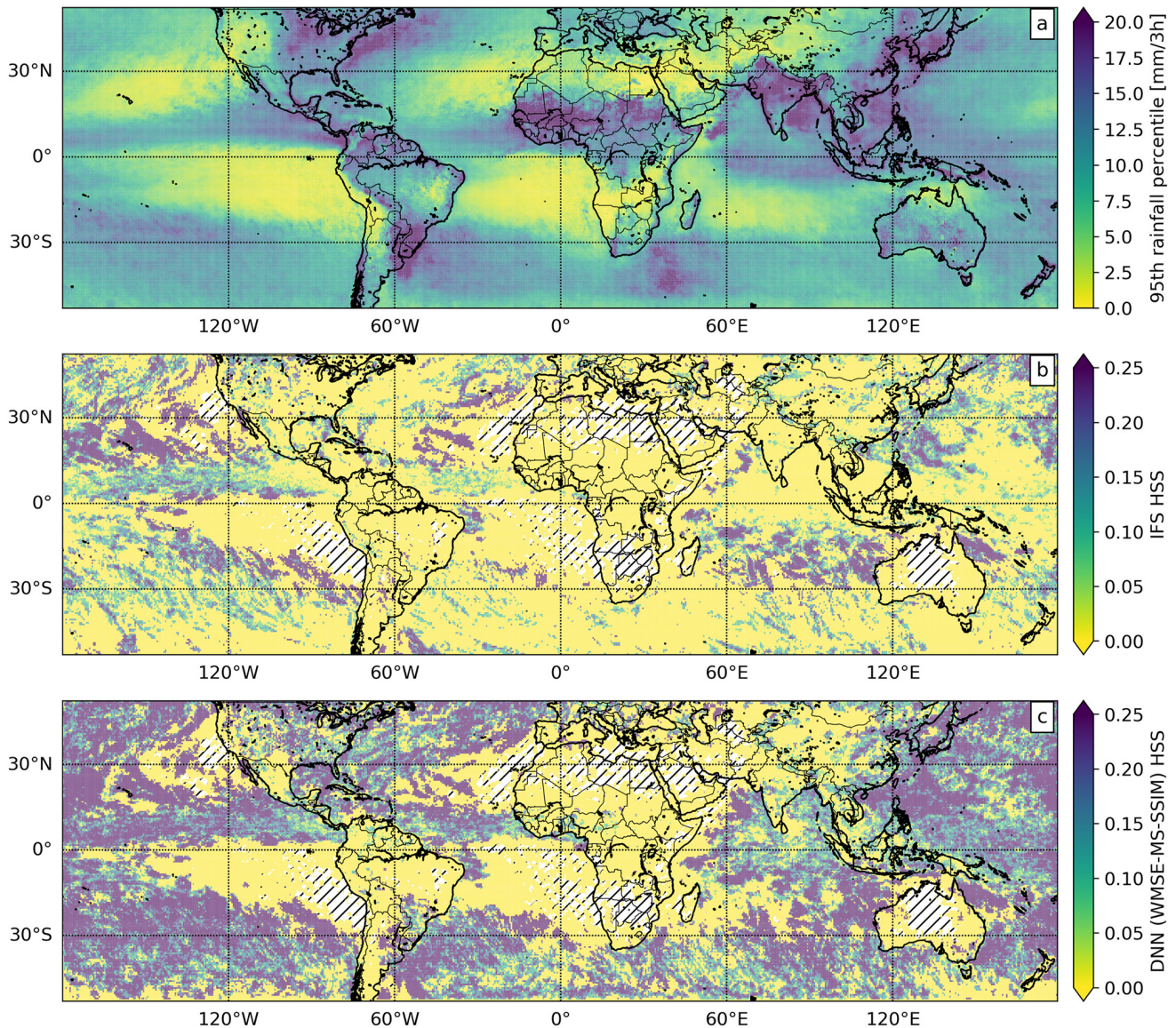
**Table 2**

*Event-Based Forecast Skill Scores for Rainfall Events Above the 95th Percentile*

Model	Loss	HSS	%	F1	%	CSI	%	POD	%	FAR	%
IFS	-	0.067	-	0.069	-	0.036	-	0.041	-	0.778	-
Quantile map.	-	0.156	133	0.161	135	0.088	144.4	0.163	299	0.840	-8
Ridge Regr.	MSE	0.040	-40	0.041	-41	0.021	-42	0.022	-46	0.775	0
DNN	MSE	0.113	69	0.115	67	0.061	69	0.066	61	<b>0.567</b>	<b>27</b>
DNN	MS-SSIM	0.174	160	0.177	157	0.097	169	0.115	180	0.622	20
DNN	WMSE-MS-SSIM	<b>0.192</b>	<b>187</b>	<b>0.195</b>	<b>183</b>	<b>0.108</b>	<b>200</b>	<b>0.139</b>	<b>239</b>	0.673	13

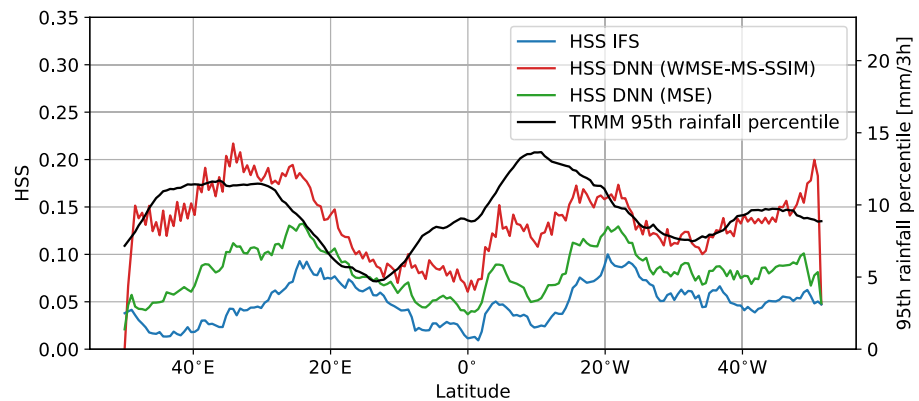
*Note.* The percentage columns give the relative improvement over the IFS mean for each error metric and skill score.





**Figure 3.** Spatial distribution of the 95th rainfall percentile and Heidke Skill Score (HSS) for events above the 95th percentile. (a) The 95th percentile of the rainfall distribution at each grid cell of the Tropical Rainfall Measurement Mission data set. (b) The spatially resolved HSS for the IFS mean. (c) The spatially resolved HSS for the deep neural network post-processed forecast, trained with the proposed WMSE-MS-SSIM loss. Hatched areas indicate grid-cells where the HSS could not be evaluated. This is due to the low number of wet times in these locations, so that the percentile thresholds could not be determined.

MS-SSIM alone as loss shows the highest scores below the 95th threshold. The proposed WMSE-MS-SSIM loss leads to significant improvements even above the 95th percentile (improving the IFS forecast by 192% in terms of the HSS) and yields the most skillfull forecast for events above the 99th percentile (improving the IFS forecast by more than 500% in terms of the HSS). Note that the FAR score is not as strongly improved as the other skills, indicating slightly more frequent false alarms when optimizing with the WMSE-MS-SSIM loss. We attribute this to the highly localized, intermittent nature of heavy rainfall events and emphasize that - in view of the results for the other error metrics - the increased number of false positives is more than balanced by the increased number of true positives. The DNN trained with the WMSE-MS-SSIM loss introduced above leads to substantial improvements also for the spatial patterns of heavy rainfall events. In particular for regions with stronger heavy rainfall events (Figure 3) the skill improvement increases. This is also visible in Figure 4 showing longitudinal averages of the 95th rainfall percentile and the HSS scores of the IFS and the DNNs trained with MSE and the WMSE-MS-SSIM loss function. There remain regions, however, where the HSS is not substantially improved. These are



**Figure 4.** Zonal mean of the Heidke Skill Score for events above the 95th percentile for the Integrated Forecast System mean (blue), the deep neural network (DNNs) trained with the mean squared error (MSE) (green) and WMSE-MS-SSIM loss (red). The zonal mean of the Tropical Rainfall Measurement Mission (TRMM) 95th rainfall percentiles are shown in black. *Note.* That the averaged HSS of the DNN (WMSE-MS-SSIM) is approximately proportional to the 95th rainfall percentiles of TRMM.

mainly given by areas for which the IFS itself already has particularly low forecast skill (Figure S7 in Supporting Information S1), including a large fraction of the land masses between 30°S and 30°N. The rainfall frequencies are still improved in this region (Figure S8 in Supporting Information S1), although less strongly than over the entire global domain.

#### 4. Discussion

We introduced a DNN to model heavy rainfall from short-range numerical weather ensemble forecasts. To address the strong statistical imbalance of the training data, a loss function is introduced that combines a weighted MSE with a structural similarity measure (WMSE-MS-SSIM). The proposed WMSE-MS-SSIM loss function is found to substantially improve the training with respect to high values compared to using the MSE and MS-SSIM individually, which are two commonly used loss functions. For comparison, we show that post-processing the IFS mean with a ridge regression model does not lead to any improvements. This motivates the importance of a non-linear DNN architecture such as the U-Net. Moreover, our results suggest that the U-Net architecture is indeed capable of capturing the multi-scale spatial structure of rainfall accurately.

The WMSE-MS-SSIM loss substantially improves relative rainfall frequencies in the DNN output, the ME and CW-SSIM of overall rainfall fields, as well as categorical skill scores for heavy rainfall events above the 90th and higher percentile, with strongly increasing relative rate of improvement for higher thresholds. As seen in Figure 3 and Figure 4 the skill improvement follows largely regions with higher rainfall percentiles. A possible explanation could be that the WMSE-MS-SSIM loss is particularly successful at locations with high rainfall values. This is supported by the seemingly lower correlation of the MSE-trained DNN's HSS with the rainfall percentiles as shown in Figure 4. In regions, where the IFS predictions are not much better than a random forecast (Heidke skill score close to 0), our DNN-based post-processing that uses these IFS prediction as input is not able to improve these IFS forecasts significantly. A direction for future research could thus be the improvement of our method in tropical regions where the skill is lower than in higher latitudes.

As noted by several authors, a single metric that captures all characteristics of a forecast does not exist, which renders the evaluation of purely data-driven weather forecasts particularly challenging (Ebert-Uphoff & Hilburn, 2020; Rasp & Thuerey, 2021; Ravuri et al., 2021). In particular, physical consistency, that is often assumed in established metrics, is not always guaranteed in neural network-based predictions. In this study, the DNN performance was manually evaluated using several metrics, both continuous and categorical. We believe that the development of more suitable and comprehensive evaluation metrics, or combinations thereof, will be an important direction of future research. It would enable a fully automatic hyperparameter tuning with respect to the various forecast qualities, which is, however, outside the scope of this study.

Taking the mean of the IFS ensemble is expected to damp the high rainfall values in the forecast. Hence, the results of the IFS shown here do not represent the skill of single ensemble members to forecast heavy rainfall events, which do not show this damping. The comparison of the bias correction methods presented in this study to the IFS ensemble mean therefore aim to show the respective relative improvements. Nevertheless, our results demonstrate the ability of the proposed DNN architecture to learn high rainfall values that are not produced by the existing precipitation parameterization of the IFS model, and to substantially improve their prediction.

The satellite-based TRMM rainfall data is used in this study as a ground truth. However, since different observational rainfall datasets usually agree only on much larger spatial and time scales than considered in this study, this should not be taken literally. The high resolution chosen for this study is important to capture the intermittent variability of heavy rainfall events. Since our method can be retrained in a flexible manner, it is possible to re-calibrate it to other observational data set as well. This allows for a continuous update of the DNN once more accurate observational datasets become available.

Interestingly, the error statistics did not change significantly when rainfall was excluded and only the vertical wind speed were considered as input features. This indicates that the DNN can learn a good representation of rainfall and especially its high values from the vertical velocity alone. This is also in agreement with previous works (Müller et al., 2020; O’Gorman & Schneider, 2009) on the link of the vertical velocity to heavy rainfall events.

An improved structural similarity in terms of the CW-SSIM is achieved when using the WMSE-MS-SSIM loss, compared to using the MS-SSIM alone as loss function. Adding the weighted MSE to the MS-SSIM loss might not be expected to increase the overall structural similarity of the DNN output. We speculate that the increased structural similarity we found might be related to the DNNs ability to improve the standard deviation that is measured in the CW-SSIM and to perform a transformation of the rainfall distribution similar to the QM method. Both our DNN and QM show accurate frequency distributions as well as relatively high structural similarity compared to the other models. Nevertheless, when trained using the WMSE-MS-SSIM loss, the forecast skill of our DNN-based post processing outperforms all other methods including QM, for almost all continuous and event-based validation metrics (see Tables 1 and 2).

A qualitative comparison of the DNN output with the TRMM target (Figure S1 in Supporting Information S1) shows that there remain small-scale features that are not captured by our method. This lack of high-frequency details in the output can be attributed to the deterministic nature of our neural network model. Here, generative models that learn stochastic functions and are therefore able to produce realistic small-scale features might offer a direction toward further improvements. However, producing stochastic small-scale features does not necessarily lead to a better forecast skill of the model, in particular for high rainfall events (Ravuri et al., 2021). We therefore believe that the results presented here could also be relevant for such stochastic approaches.

Although the considered forecast has a high temporal resolution of three hours, the forecast lead time of up to twelve hours is still comparably short. With applications to disaster prevention in mind, an extension of the study to longer forecast lead times will be an important direction for future research.

With ongoing global warming, the characteristics of heavy rainfall events are expected to change. To account for this non-stationarity, the training of the proposed method can in principle be continued over time when new training data becomes available. Further, making use of the entire IFS ensemble will allow to incorporate uncertainties into the framework, which are essential for operational forecasting of heavy rainfall events.

### Data Availability Statement

Data pre-processing was done using the Climate Data Operator (CDO) software (Schulzweida, 2019) for regriding as well as the Xarray 0.15.1 package. The Pytorch 1.7.0 (Paszke et al., 2019) source code for training and data processing will be available at (<https://zenodo.org/badge/latestdoi/457716105>) on publication. The QM method was implemented using the Xclim 0.25.0 Python package (Logan et al., 2021). The IFS training data is available for download at the Copernicus Climate Change Service (C3S; Hersbach et al., 2020; <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels>). The TRMM (TMPA) data can be obtained at the Goddard Earth Sciences Data and Information Services Center (GES DISC; Leptoukh, 2005; [https://disc.gsfc.nasa.gov/datasets/TRMM\\_3B42\\_7/summary](https://disc.gsfc.nasa.gov/datasets/TRMM_3B42_7/summary)).

**Acknowledgments**

The authors would like to thank the three anonymous reviewers for their helpful comments and suggestions. The authors acknowledge funding by the Volkswagen Foundation, as well as the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research and the Land Brandenburg for supporting this project by providing resources on the high performance computer system at the Potsdam Institute for Climate Impact Research. Open access funding enabled and organized by Projekt DEAL.

**References**

Agrawal, S., Barrington, L., Bromberg, C., Burge, J., Gazen, C., & Hickey, J. (2019). *Machine learning for precipitation nowcasting from radar images*.

Ali, H., Fowler, H. J., & Mishra, V. (2018). Global observational evidence of strong linkage between dew point temperature and precipitation extremes. *Geophysical Research Letters*, *45*(22), 12–320. <https://doi.org/10.1029/2018gl080557>

Allan, R. P., & Soden, B. J. (2008). Atmospheric warming and the amplification of precipitation extremes. *Science*, *321*(5895), 1481–1484. <https://doi.org/10.1126/science.1160787>

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. <https://doi.org/10.1038/nature14956>

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I., et al. (2019). Mswep v2 global 3-hourly 0.1 precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, *100*(3), 473–500. <https://doi.org/10.1175/bams-d-17-0138.1>

Berg, P., Feldmann, H., & Panitz, H.-J. (2012). Bias correction of high resolution regional climate model data. *Journal of Hydrology*, *448*, 80–92. <https://doi.org/10.1016/j.jhydrol.2012.04.026>

Bocquet, M., Farchi, A., & Malartic, Q. (2020). Online learning of both state and dynamics using ensemble kalman filters. *Foundations of Data Science*, *3*(3), 305. <https://doi.org/10.3934/fods.2020015>

Boers, N., Bookhagen, B., Marengo, J., Marwan, N., vonStorch, J.-S., & Kurths, J. (2015). Extreme rainfall of the south American monsoon system: A dataset comparison using complex networks. *Journal of Climate*, *28*(3), 1031–1056. <https://doi.org/10.1175/jcli-d-14-00340.1>

Boyle, J., & Klein, S. A. (2010). Impact of horizontal resolution on climate model forecasts of tropical precipitation and diabatic heating for the twp-ice period. *Journal of Geophysical Research: Atmospheres*, *115*, D23113. <https://doi.org/10.1029/2010jd014262>

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2728–2744. <https://doi.org/10.1029/2019ms001711>

Cannon, A. J., Sobie, S. R., & Murdock, T. Q. (2015). Bias correction of gcm precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *Journal of Climate*, *28*(17), 6938–6959. <https://doi.org/10.1175/jcli-d-14-00754.1>

Déqué, M. (2007). Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global and Planetary Change*, *57*(1–2), 16–26. <https://doi.org/10.1016/j.gloplacha.2006.11.030>

Donat, M., Alexander, L., Yang, H., Durre, I., Vose, R., Dunn, R., et al. (2013). Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The hadex2 dataset. *Journal of Geophysical Research: Atmospheres*, *118*(5), 2098–2118. <https://doi.org/10.1002/jgrd.50150>

Ebert-Uphoff, I., & Hilburn, K. (2020). Evaluation, tuning, and interpretation of neural networks for working with images in meteorological applications. *Bulletin of the American Meteorological Society*, *101*(12), E2149–E2170. <https://doi.org/10.1175/bams-d-20-0097.1>

European Centre for Medium-Range Weather Forecasts. (2012). *The ECMWF ensemble prediction system*. Retrieved from [https://www.ecmwf.int/sites/default/files/the\\_ECMWF\\_Ensemble\\_prediction\\_system.pdf](https://www.ecmwf.int/sites/default/files/the_ECMWF_Ensemble_prediction_system.pdf)

Fischer, E. M., & Knutti, R. (2016). Observed heavy precipitation increase confirms theory and early models. *Nature Climate Change*, *6*(11), 986–991. <https://doi.org/10.1038/nclimate3110>

Franch, G., Nerini, D., Pendesini, M., Coviello, L., Jurman, G., & Furlanello, C. (2020). Precipitation nowcasting with orographic enhanced stacked generalization: Improving deep learning predictions on extreme events. *Atmosphere*, *11*(3), 267. <https://doi.org/10.3390/atmos11030267>

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2021). Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200092. <https://doi.org/10.1098/rsta.2020.0092>

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., & Engen-Skaugen, T. (2012). Downscaling rcm precipitation to the station scale using statistical transformations—a comparison of methods. *Hydrology and Earth System Sciences*, *16*(9), 3383–3390. <https://doi.org/10.5194/hess-16-3383-2012>

Guerreiro, S. B., Fowler, H. J., Barbero, R., Westra, S., Lenderink, G., Blenkinsop, S., et al. (2018). Detection of continental-scale intensification of hourly rainfall extremes. *Nature Climate Change*, *8*(9), 803–807. <https://doi.org/10.1038/s41558-018-0245-3>

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*, 1999–2049.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. <https://doi.org/10.1080/00401706.1970.10488634>

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., et al. (2007). The trmm multisatellite precipitation analysis (tmap): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, *8*(1), 38–55. <https://doi.org/10.1175/jhm560.1>

Kingma, D. P., & Ba, J. (2017). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.

Koutsoyiannis, D. (2004b). Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation. *Hydrological Sciences Journal*, *49*(4), 575–590. <https://doi.org/10.1623/hysj.49.4.575.54430>

Koutsoyiannis, D. (2004a). Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records. *Hydrological Sciences Journal*, *49*(4), 591–610. <https://doi.org/10.1623/hysj.49.4.591.54424>

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. <https://doi.org/10.1038/nature14539>

Leptoukh, G. (2005). Nasa remote sensing data in Earth sciences: Processing, archiving, distribution, applications at the ges disc. In *Proceedings of the 31st intl symposium of remote sensing of environment*.

Logan, T., Bourgault, P., Smith, T. J., Huard, D., Biner, S., Labonté, M.-P., et al. (2021). Ouranosinc/xclim: V0.31.0. Zenodo. [Data set]. <https://doi.org/10.5281/zenodo.5649661>

Maraun, D. (2016). Bias correcting climate change simulations—a critical review. *Current Climate Change Reports*, *2*(4), 211–220. <https://doi.org/10.1007/s40641-016-0050-x>

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., et al. (2021). *IPCC, 2021: Climate change 2021: The physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*. Cambridge University Press. In Press.

Müller, A., Niedrich, B., & Névir, P. (2020). Three-dimensional potential vorticity structures for extreme precipitation events on the convective scale. *Tellus A: Dynamic Meteorology and Oceanography*, *72*(1), 1–20. <https://doi.org/10.1080/16000870.2020.1811535>

O’Gorman, P. A., & Schneider, T. (2009). The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. *Proceedings of the National Academy of Sciences*, *106*(35), 14773–14777. <https://doi.org/10.1073/pnas.0907610106>

- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, pp. 8024–8035). Curran Associates, Inc.
- Pfahl, S., O'Gorman, P. A., & Fischer, E. M. (2017). Understanding the regional pattern of projected future changes in extreme precipitation. *Nature Climate Change*, 7(6), 423–427. <https://doi.org/10.1038/nclimate3287>
- Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11), e2020MS002203. <https://doi.org/10.1029/2020ms002203>
- Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11), 3885–3900. <https://doi.org/10.1175/mwr-d-18-0187.1>
- Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39), 9684–9689. <https://doi.org/10.1073/pnas.1810286115>
- Rasp, S., & Thuerey, N. (2021). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, 13(2), e2020MS002405. <https://doi.org/10.1029/2020ms002405>
- Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., et al. (2021). Skillful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878), 672–677. <https://doi.org/10.1038/s41586-021-03854-z>
- Ronneberger, O., Fischer, P., & Brox, T. (2015). In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer. U-net: Convolutional networks for biomedical image segmentation.
- Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C., & Markey, M. K. (2009). Complex wavelet structural similarity: A new image similarity index. *IEEE Transactions on Image Processing*, 18(11), 2385–2401. <https://doi.org/10.1109/tip.2009.2025923>
- Schulzweida, U. (2019). Cdo user guide. (Version 1.9. 6). *Max Planck Institute for Meteorology*, 53, 20146. <https://doi.org/10.5281/zenodo.3539275>
- Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., kin Wong, W., & chun Woo, W. (2017). Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in Neural Information Processing Systems*, 30.
- Sønderby, C. K., Espelholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., et al. (2020). *Metnet: A neural weather model for precipitation forecasting*. arXiv preprint arXiv:2003.12140.
- Tong, Y., Gao, X., Han, Z., Xu, Y., Xu, Y., & Giorgi, F. (2021). Bias correction of temperature and precipitation over China for rcm simulations using the qm and qdm methods. *Climate Dynamics*, 57(5), 1425–1443. <https://doi.org/10.1007/s00382-020-05447-4>
- Tran, Q.-K., & Song, S.-k. (2019). Computer vision in precipitation nowcasting: Applying image quality assessment metrics for training deep neural networks. *Atmosphere*, 10(5), 244. <https://doi.org/10.3390/atmos10050244>
- Traxl, D., Boers, N., Rheinwalt, A., & Bookhagen, B. (2021). The role of cyclonic activity in tropical temperature-rainfall scaling. *Nature Communications*, 12(1), 1–9. <https://doi.org/10.1038/s41467-021-27111-z>
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612. <https://doi.org/10.1109/tip.2003.819861>
- Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). In *The thirty-seventh asilomar conference on signals* (Vol. 2, pp. 1398–1402). systems & computers. Multiscale structural similarity for image quality assessment.
- Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J. J., et al. (2021). Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, 48(15), e2021GL092555. <https://doi.org/10.1029/2021GL092555>
- Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9), e2020MS002109. <https://doi.org/10.1029/2020ms002109>
- Wilcox, E. M., & Donner, L. J. (2007). The frequency of extreme rain events in satellite rain-rate estimates and an atmospheric general circulation model. *Journal of Climate*, 20(1), 53–69. <https://doi.org/10.1175/jcli3987.1>
- Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences* (Vol. 100). Academic press.
- Xu, C.-y. (1999). From gcms to river flow: A review of downscaling methods and hydrologic modelling approaches. *Progress in Physical Geography*, 23(2), 229–249. <https://doi.org/10.1177/030913339902300204>
- Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, 11(1), 1–10. <https://doi.org/10.1038/s41467-020-17142-3>

# Supporting Information for ”Deep Learning for Improving Numerical Weather Prediction of Heavy Rainfall”

Philipp Hess<sup>1,2</sup>, Niklas Boers<sup>1,2,3</sup>

<sup>1</sup>Technical University of Munich, Germany; School of Engineering & Design, Earth System Modelling

<sup>2</sup>Potsdam Institute for Climate Impact Research (PIK), Telegraphenberg A31, Potsdam, 14473, Germany

<sup>3</sup>Department of Mathematics and Global Systems Institute, University of Exeter, Exeter, UK

## Contents of this file

1. Text S1 to S2
2. Figures S1 to S8
3. Table S1

---

Corresponding author: Philipp Hess, philipp.hess@tum.de

**Text S1.** The root mean square error (RMSE) and mean error (ME) are defined as,

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2}, \quad (1)$$

$$\text{ME} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i), \quad (2)$$

where  $N$  is the number of training examples,  $y$  is the TRMM target and  $\hat{y}$  is the modelled rainfall output. The multi-scale structural similarity measure (MS-SSIM)(Wang et al., 2003) quantifies the structural similarity between two images, in our case two spatial rainfall maps, as sets of  $N$  grid-cells, i.e.  $\mathbf{y} = \{y_i | i = 1, 2, \dots, N\}$  and  $\hat{\mathbf{y}} = \{\hat{y}_i | i = 1, 2, \dots, N\}$ . The MS-SSIM then iteratively computes three measures, for luminance  $l(\mathbf{y}, \hat{\mathbf{y}})$ , contrast  $c(\mathbf{y}, \hat{\mathbf{y}})$  and structure  $s(\mathbf{y}, \hat{\mathbf{y}})$  by successively downsampling and low-pass filtering the image signals. The three measures are defined as

$$l(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2\mu_y\mu_{\hat{y}} + C_1}{\mu_y^2 + \mu_{\hat{y}}^2 + C_1}, \quad (3)$$

$$c(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2\sigma_y\sigma_{\hat{y}} + C_2}{\sigma_y^2 + \sigma_{\hat{y}}^2 + C_2}, \quad (4)$$

$$s(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\sigma_{y\hat{y}} + C_3}{\sigma_y\sigma_{\hat{y}} + C_3}, \quad (5)$$

where  $\mu_y$  is the mean,  $\sigma_y$  the standard deviation of  $\mathbf{y}$  and  $\sigma_{y,\hat{y}}$  the covariance of  $\mathbf{y}$  and  $\hat{\mathbf{y}}$ . The small constants  $C_1$ ,  $C_2$ , and  $C_3$  are included to improve the stability and are computed via

$$C_1 = (K_1L)^2, C_2 = (K_2L)^2 \text{ and } C_3 = C_2/2, \quad (6)$$

where  $L = 255$ ,  $K_1 = 0.01$  and  $K_2 = 0.03$ . The MS-SSIM can then be written as,

$$\text{MS-SSIM}(\mathbf{y}, \hat{\mathbf{y}}) = [l_M(\mathbf{y}, \hat{\mathbf{y}})]^{\alpha_M} \cdot \prod_{j=1}^M [c_j(\mathbf{y}, \hat{\mathbf{y}})]^{\beta_j} \cdot [s_j(\mathbf{y}, \hat{\mathbf{y}})]^{\gamma_j}, \quad (7)$$

where  $M$  denotes the number downsampling iterations. The exponents  $\alpha_M$ ,  $\beta_j$  and  $\gamma_j$  can be adjusted to give different weights to the measures, but are set to  $\alpha_j = \beta_j = \gamma_j = 1$ .

The complex wavelet structural similarity (CW-SSIM)(Sampat et al., 2009), extends the idea of structural similarity to the complex wavelet domain. The motivation behind it is that structural changes between two images, such as small rotations or translations will lead to a constant relative phase shift in the coefficients of a complex wavelet transform. Therefore, the CW-SSIM is constructed in such a way that it is insensitive to *relative* phase shifts and magnitude distortions. On the other hand it is sensitive to non-structural transformations in images, such as changes in sharpness, that will lead to phase shifts in the coefficients. The CW-SSIM is defined as

$$\text{CW-SSIM}(\mathbf{c}_y, \mathbf{c}_{\hat{y}}) = \frac{2|\sum_{i=1}^N c_{y,i}c_{\hat{y},i}^*| + C}{\sum_{i=1}^N |c_{y,i}|^2 + \sum_{i=1}^N |c_{\hat{y},i}|^2 + C}, \quad (8)$$

where  $\mathbf{c}_y = \{c_{y,i}|i = 1, 2, \dots, N\}$  and  $\mathbf{c}_{\hat{y}} = \{c_{\hat{y},i}|i = 1, 2, \dots, N\}$  are two sets of complex wavelet coefficients obtained at the same spatial location and wavelet subbands of the two images being compared. The asterix denotes the complex conjugate and  $C = 0.01$  is a small constant for stability.

**Text S2.** We quantify the forecast skill of extreme events with categorical skill scores commonly used in meteorology and machine learning, such as the critical success index (CSI), probability of detection (POD), false alarm ratio (FAR), F1 and Heidke skill score (HSS). These skill scores can be computed from the contingency table (see Table S1). The table classifies event forecast outcomes into true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). Based on these categories, the skill scores can be defined as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$



$$F1 = 2 \frac{\text{Recall Precision}}{\text{Recall} + \text{Precision}},$$

$$HSS = \frac{2(TP \cdot TN - FP \cdot FN)}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)},$$

$$CSI = \frac{TP}{TP + FN + FP},$$

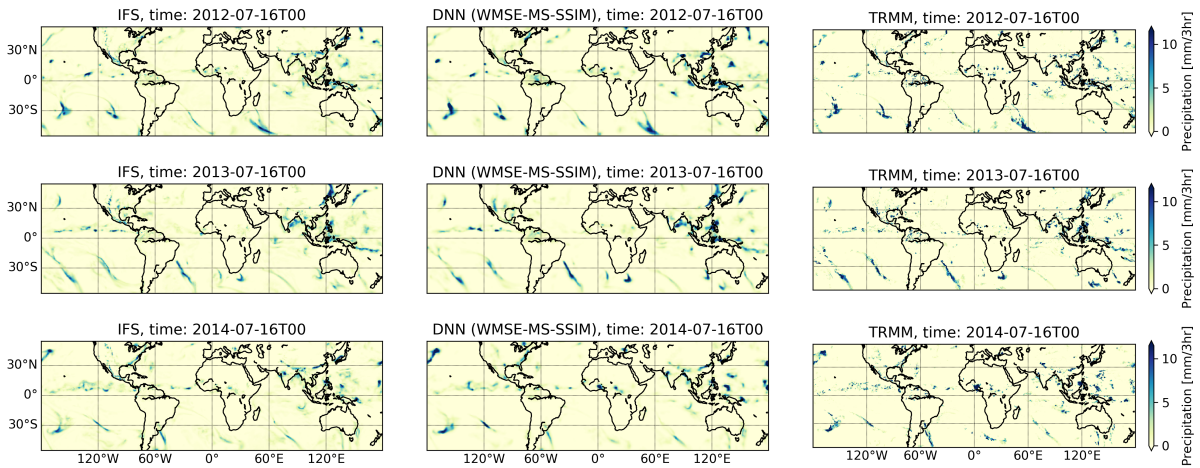
$$POD = \text{Recall},$$

$$FAR = \frac{FP}{FP + TP}.$$

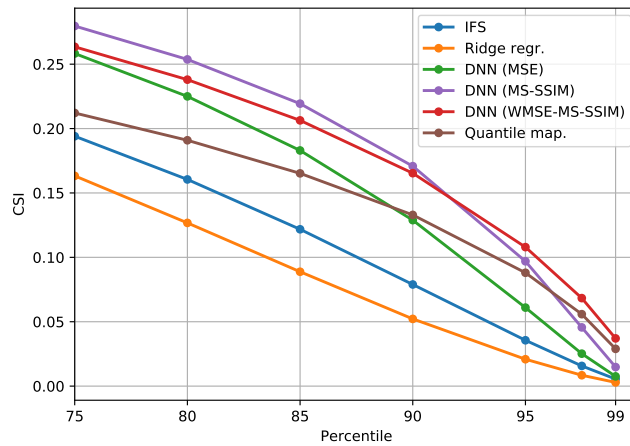
The recall score computes the proportion of relevant events that were classified correctly and precision gives the fraction of positive classifications that were correct. The F1 score combines precision and recall as a harmonic mean and is commonly used in machine learning to evaluate predictions on strongly imbalanced data. The Heidke Skill Score (HSS) evaluates the accuracy of event predictions, e.g. rainfall extremes, relative to a random forecast and can also be used for strongly imbalanced classes. The critical success (CSI) relates the accuracy of event predictions to the actually observed events, without accounting for correct negative predictions. The probability of detection (POD) and false alarm ratio (FAR) scores should be assessed together, where the former is defined identically to the recall score. Since POD ignores false alarms, the false alarms ratio (FAR) can be used to evaluate these.

## References

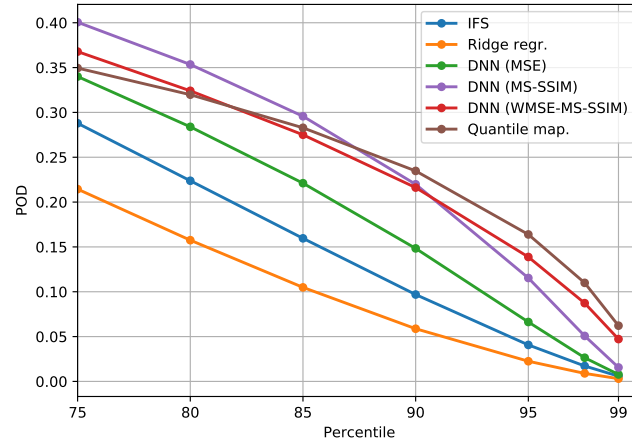
- Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C., & Markey, M. K. (2009). Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, 18(11), 2385–2401.
- Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The thirty-seventh asilomar conference on signals, systems & computers, 2003* (Vol. 2, pp. 1398–1402).



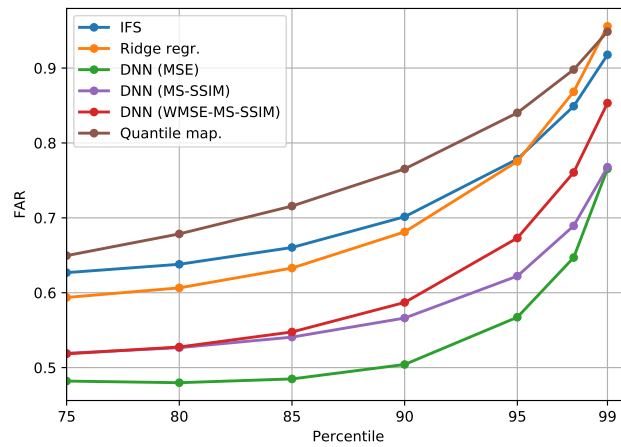
**Figure S1.** Spatial rainfall maps at three different time instances of the IFS, the DNN and TRMM test set data are shown for a qualitative comparison.



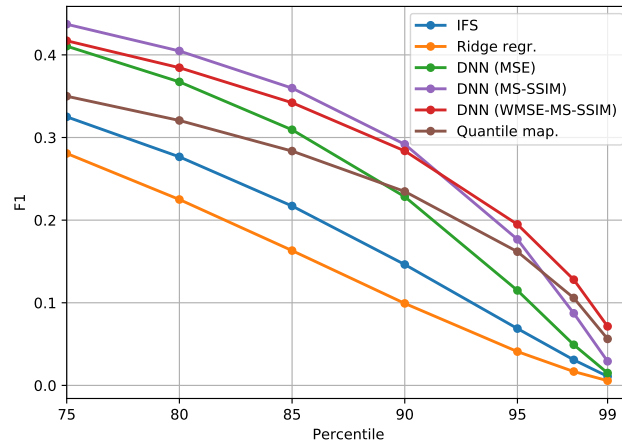
**Figure S2.** The critical success index (CSI) for rainfall events above the 75th percentile threshold.



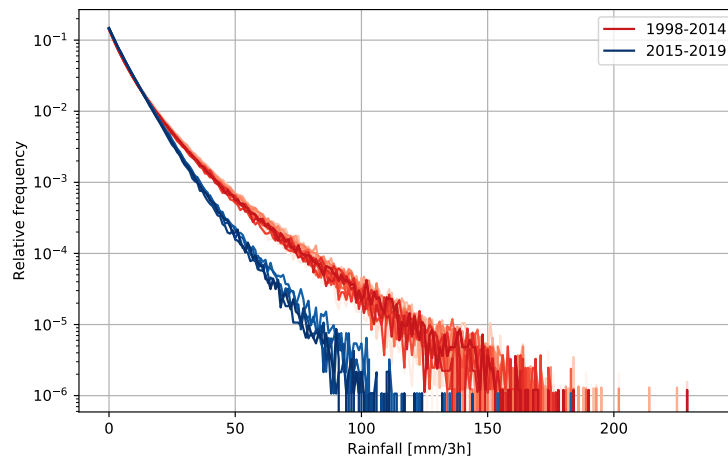
**Figure S3.** The probability of detection (POD) of rainfall events above the 75th percentile threshold.



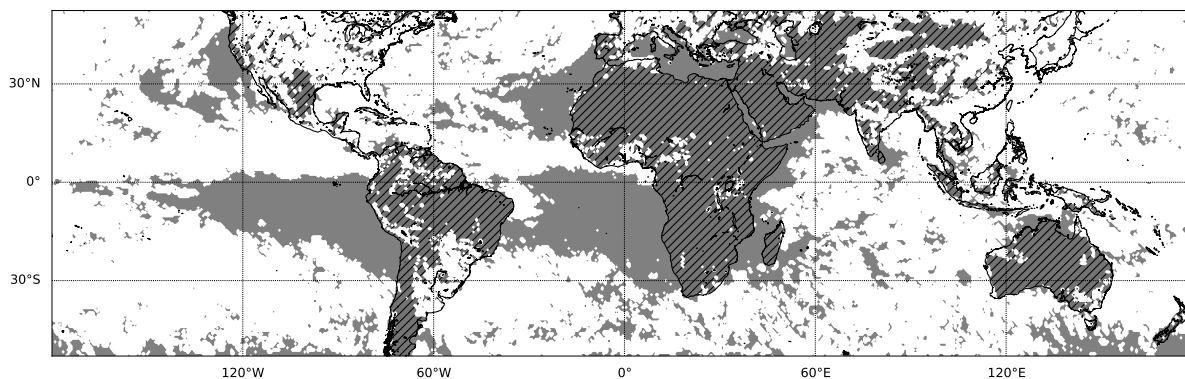
**Figure S4.** The false alarm ratio (FAR) of rainfall events above the 75th percentile threshold.



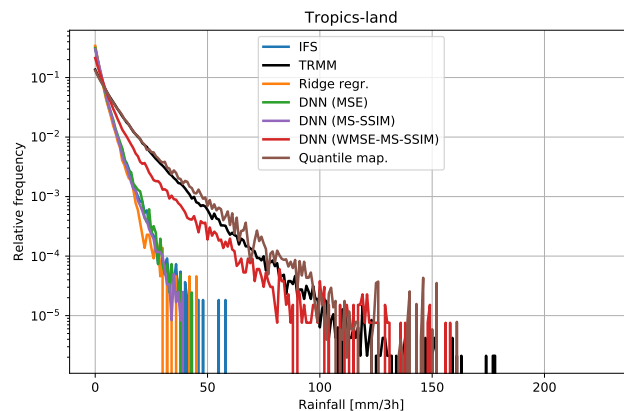
**Figure S5.** The F1 score for rainfall events above the 75th percentile threshold.



**Figure S6.** The histograms of grid-cell values show here are computed over the entire part of the globe covered by the TRMM data ( $50^{\circ}\text{S}$  to  $50^{\circ}\text{N}$ ) and for single years. The histograms of years before 2015 are colored in red and for years thereafter in blue.



**Figure S7.** Intersection of the Heidke skill score below 0.01 of the IFS and the DNN (WMSE-MS-SSIM) forecasts, shown in grey. Land area is hatched for better distinction. In the highlighted regions, the DNN does not show significant improvements over the relatively low IFS forecast skill. In the tropics (30°S to 30°N) this includes a large fraction of the land area, e.g. parts of South America, Africa and Australia.



**Figure S8.** Histograms of three-hourly rainfall event magnitudes over land in the tropics (30°S to 30°N) are shown on a logarithmic y-axis for TRMM (black), IFS (blue), ridge regression (orange), DNN trained with the MSE loss (green), the MS-SSIM loss (purple) and the WMSE-MS-SSIM loss (red)

**Table S1.** Contingency table of forecast outcomes for binary events.

	Observed	Not observed
Forecasted	True positive (TP)	False positive (FP)
Not forecasted	False negative (FN)	True negative (TN)





### 3 Generative models for improving precipitation fields from climate simulations

*Based on*

Physically constrained generative adversarial networks  
for improving precipitation fields from  
Earth system models

**Philipp Hess**, Markus Drüke, Stefan Petri,  
Felix M. Strnad and Niklas Boers, 2022.  
Nature Machine Intelligence, 4(10), 828-839.

**Copyright** The article published in Nature Machine Intelligence is published closed-access, with Springer Nature or its licensor holding exclusive rights to this article under a publishing agreement with the authors.

**Contribution** PH and NB conceived the research and designed the study with input from all authors. PH performed the data processing, neural network training, and analysis. MD conducted the CM2Mc-LPJmL experiments. All authors interpreted and discussed the results. PH wrote the manuscript with input from all authors.

*and*

Deep Learning for bias-correcting comprehensive  
high-resolution Earth system models.

**Philipp Hess**, Stefan Lange and Niklas Boers, 2022.  
Submitted to Proceedings of the National Academy of Sciences.

**Copyright** The study has not been published yet. A preprint under the [Creative Commons Attribution-NonCommercial-NoDerivatives](#) license is available on arXiv.

**Contribution** PH and NB conceived the research and designed the study with input from SL. PH performed the data processing, neural network training and analysis. All authors interpreted and discussed the results. PH wrote the manuscript with input from SL and NB.

**Summary** To simulate the Earth system over decades to millennia, numerical Earth system models have a relatively coarse horizontal resolution of around 25–200km and thus require parameterizations of small-scale processes that cannot be resolved explicitly. These approximations and limited model complexity lead to biases in temporal distributions and spatial patterns, the latter often being overly smooth compared to observations.

Impact models are developed and calibrated with observations-based data products but use the Earth system simulation output as input for future climate scenarios. This can lead to inconsistencies since the impact model is forced with input data that has not been calibrated.

Therefore, bias-correction methods aim to close this gap by correcting systematic errors in the simulations using historical observations. Established methods typically based on quantile mapping (QM) ([Cannon et al., 2015](#)) correct the local distributions in time for each grid cell individually.

The two studies P2 and P3, in this chapter tackle the bias correction problem as an unpaired image-to-image translation task, or “style transfer”, where image characteristics (such as of paintings, photos, etc.) are translated onto another image without altering the overall content. Using convolutional neural networks that can leverage spatial correlations for image translation, spatial patterns can be corrected. This is not possible with established QM-based methods, which only transform statistics locally at each grid cell without any spatial context.

The studies show how cycle-consistent generative adversarial networks (CycleGAN) ([Zhu et al., 2017](#)) can be applied as a novel bias-correction method for ESM simulations. The first study uses efficient low-resolution simulations of the CM2Mc-LPJmL ESM ([Drüke et al., 2021](#)) with reduced model complexity. It demonstrates how both temporal distribution and spatial patterns can be corrected, thus creating realistic climate simulations of precipitation that are computationally very efficient. It further investigates the generalization of the GAN to out-of-sample tasks, i.e., the extrapolation to predictions of climate states that are not seen during the network training. We introduce an architecture constraint that preserves the global ESM precipitation sum by rescaling it. This enables the GAN to capture the transient trends expected in an extreme global warming scenario. Additionally, we apply an interpretability method ([Smilkov et al., 2017](#)) to the GAN architecture. Using the gradients of the discriminator network, we find that the model has learned a physically consistent distinction between generated and real-world

precipitation fields. The second study extends the approach to high-resolution and comprehensive ESM simulations and evaluates the method against the state-of-the-art bias-correction framework ISIMIP3BASD (Lange, 2019). It shows that even the most advanced ESMs can still be significantly improved with the GAN post-processing, particularly the characteristic intermittency of spatial patterns. The strong ISIMIP3BASD baseline is outperformed on the tasks of improving spatial patterns and compares similarly on temporal distributions. It can be further combined with our method to achieve the best overall results.

## **3.1 P2 | Physically constrained generative adversarial networks for improving precipitation fields from Earth system models**

Please turn to the next page.

# Physically Constrained Generative Adversarial Networks for Improving Precipitation Fields from Earth System Models

Philipp Hess<sup>1,2</sup>, Markus Drüke<sup>2</sup>, Stefan Petri<sup>2</sup>, Felix M. Strnad<sup>2,3</sup>, and Niklas Boers<sup>1,2,4</sup>

<sup>1</sup>Technical University Munich, Munich, Germany; School of Engineering & Design, Earth System Modelling

<sup>2</sup>Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany

<sup>3</sup>Cluster of Excellence - Machine Learning for Science, Eberhard Karls Universität Tübingen, Germany

<sup>4</sup>Global Systems Institute and Department of Mathematics, University of Exeter, Exeter, UK

## Key Points:

- A generative adversarial network improves both distributions and spatial structure of the precipitation output of a numerical Earth system model.
- Constraining its architecture enables the network to generalize to transient future climates not seen during training.
- A gradient-based interpretability method shows that the network has learned to identify geographical regions with strong model biases.

---

Corresponding author: Philipp Hess, [philipp.hess@tum.de](mailto:philipp.hess@tum.de)

## Abstract

Precipitation results from complex processes across many scales, making its accurate simulation in Earth system models (ESMs) challenging. Existing post-processing methods can improve ESM simulations locally, but cannot correct errors in modelled spatial patterns. Here we propose a framework based on physically constrained generative adversarial networks (GANs) to improve local distributions and spatial structure simultaneously. We apply our approach to the computationally efficient ESM CM2Mc-LPJmL. Our method outperforms existing ones in correcting local distributions, and leads to strongly improved spatial patterns especially regarding the intermittency of daily precipitation. Notably, a double-peaked Intertropical Convergence Zone, a common problem in ESMs, is removed. Enforcing a physical constraint to preserve global precipitation sums, the GAN can generalize to future climate scenarios unseen during training. Feature attribution shows that the GAN identifies regions where the ESM exhibits strong biases. Our method constitutes a general framework for correcting ESM variables and enables realistic simulations at a fraction of the computational costs.

## 1 Introduction

Numerical Earth system models (ESMs) simulate the dynamics of Earth system components such as the atmosphere, oceans, vegetation, and polar ice-sheets, as well as their interactions, by solving the relevant partial differential equations on discretized spatial grids. The grid resolution is limited by computational costs. For state-of-the-art comprehensive ESMs, integrating the differential equations requires parallelized runs on thousands of CPU cores. The finite resolution requires processes on unresolved spatial scales to be parameterized, i.e., to be written as functions of the resolved variables. This introduces a source for potential errors in ESMs. It is generally expected that the accuracy of ESM simulations increases with increasing resolution of the spatial grid on which the model is integrated (Palmer & Stevens, 2019).

A higher grid resolution, however, comes at even higher computational cost, and trade-offs are therefore typically necessary. The time current state-of-the-art ESMs take to make projections for the decadal to centennial time scales relevant in the context of anthropogenic climate change render it challenging to simulate ensembles with sufficient size for a thorough uncertainty quantification. Similarly, the high computational cost even for simulating single trajectories prevent more systematic parameter calibration. Complementary to high-resolution but computationally demanding ESMs, efficient model setups that are still as accurate as possible are therefore also needed.

The generation of precipitation involves a wide range of physical processes, from microscopic interactions of droplets in clouds over atmospheric convection to synoptic-scale weather systems. The resulting complex dynamics needs to be captured accurately to model the high variability and intermittency of precipitation in both space and time. A reduced resolution and limited number of explicitly resolved processes in ESMs therefore leads to errors that can strongly affect the representation of sub-grid scale processes such as precipitation (Wilcox & Donner, 2007; Boyle & Klein, 2010; IPCC, 2021).

These errors can be addressed in a local or point-wise manner by applying post-processing methods to the individual simulated time series. Traditionally, this is done by relating the statistics of a historical model simulation with observations. Quantile mapping (QM), in particular, has become a popular method for improving the model output statistics of precipitation (Déqué, 2007; Tong et al., 2021; Gudmundsson et al., 2012; Cannon et al., 2015). It approximates a mapping from the estimated cumulative distribution function of the modelled to the observed quantity over a historical period. The inferred mapping can then be applied to correct new data. QM gives good results in correcting temporal distributions locally, i.e., errors in the distribution at a given grid cell. QM is, however, not able to improve the spatial structure of the modelled output, such as its intermittency

especially for the case of precipitation. For this task a spatial context larger than the single grid cells used to compute the distributions in QM is required. It should be emphasized that even a (almost) perfect reproduction of the distributions at each grid cell would by no means guarantee that also the spatial patterns are reproduced accurately. In particular, the patterns may still be too smooth and lack the spatial intermittency that is typical for realistic precipitation fields.

Machine learning (ML) methods from image-to-image translation in computer vision offer a new approach to improve the structure of ESM output in the spatial dimension. Recently, artificial neural networks have been applied successfully to post-processing tasks of numerical weather prediction and climate models (Rasp & Lerch, 2018; Grönquist et al., 2021; François et al., 2021). In weather forecasting, the trajectories of the observed state and the numerical weather model starting at an initial condition taken from observations can be directly and quantitatively compared. This allows to train discriminative ML models such as deep neural networks (LeCun et al., 2015) to directly minimize a pixel-wise distance measure as a regression task.

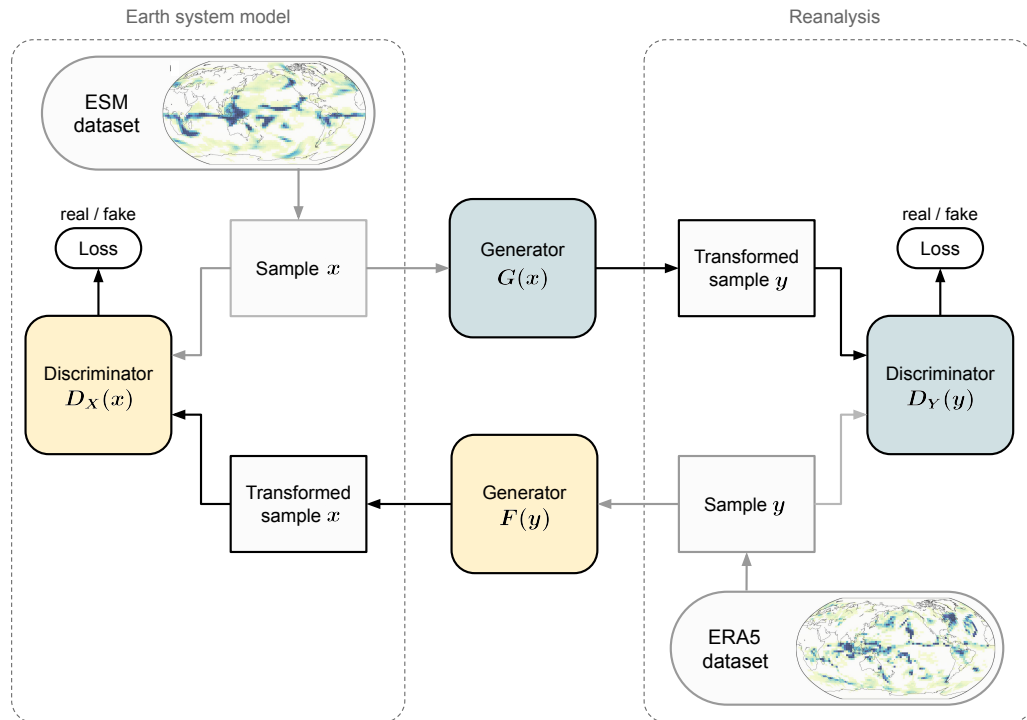
For ESMs tasked with climate projections, such a pixel-wise ground truth is not available, rendering a direct comparison between observed and modelled trajectories impossible. In particular, ML models cannot be trained via minimizing differences between simulations and corresponding observations in this case. The goal of ESMs is indeed to produce long-term summary statistics rather than to agree with observations on short time scales. In this context, generative adversarial networks (GANs) (Goodfellow et al., 2014; Mirza & Osindero, 2014; Isola et al., 2017) have emerged as suitable ML models. GANs learn to approximate a target distribution from which realistic samples can be drawn. Crucially, recent developments have shown successful application of cycle-consistent GANs (Zhu et al., 2017; Yi et al., 2017; Hoffman et al., 2018) to training tasks that do not require pairwise training samples. This suggests the suitability of cycle-consistent GANs for post-processing Earth system model simulations, for which no direct observational counterpart exists. By learning stochastic functions, GANs can also model the small-scale variability that cannot be predicted deterministically. This enables them to overcome the problem of blurring that is often found in neural network predictions (Ravuri et al., 2021). Based on these properties, GANs have been proposed for sub-grid scale parameterizations (Gagne et al., 2020) and statistical downscaling of numerical weather forecasts (Price & Rasp, 2022; L. Harris et al., 2022). Employing GANs in a post-processing task of a regional climate model, François et al. (2021) found a comparable bias correction skill of their GAN compared to quantile mapping.

Training ML algorithms typically requires the training data and separate test sets for predictions to be independent and identically distributed. When applied to historical observations and transient ESM time series with changing forcing, however, the underlying distributions are non-stationary, i.e., training and test distributions are different. In particular in the context of anthropogenic climate change, this has made the application of ML methods challenging. To generalize to such out-of-sample predictions, physics-informed or constrained neural networks have been proposed. These methods incorporate physical knowledge into the neural network through penalties in the loss function (Raissi et al., 2019), or include additional layers (Beucler et al., 2021) in the architecture.

Here, we introduce a physically constrained GAN (see Fig. 1 and Methods for details) to improve the precipitation output of ESMs, and demonstrate its performance by applying it to the CM2Mc-LPJmL model (Drüke, von Bloh, et al., 2021). We frame the post-processing as an image-to-image translation task with unpaired training samples. The first image domain corresponds to the ESM simulations, and the second to daily precipitation fields from the ERA5 reanalysis “ground truth” (Hersbach et al., 2020), spanning the period between 1950 and 2014. The translation is performed with a CycleGAN (Zhu et al., 2017), consisting of two generator-discriminator pairs, that learn bijective mappings between the ESM and reanalysis domains, with consistent translation cycles. We add a physical constraint as an

additional layer to the generator network architecture after training in order to preserve the global precipitation sum (see Methods).

We compare our results to QM-based post-processing as well as the output of a considerably more complex and higher-resolution, state-of-the-art ESM from Phase 6 of the Coupled Model Intercomparison Project (CMIP6), namely the GFDL-ESM4 (Krasting et al., 2018) model. Further, the ability of the GAN to generalize to transient future climate scenarios is evaluated for physically constrained and unconstrained GAN architectures. When applying neural network models to future projections that cannot (yet) be verified, transparency of the method becomes important. Therefore, we examine whether the GAN’s feature attribution is physically reasonable, using the SmoothGrad (Smilkov et al., 2017) interpretability method (Methods). Moreover, the quantitative interpretation of the GAN results allows us to identify regions with particularly large biases of the underlying process-based ESM, which will in turn be helpful for improving its representation of relevant physical mechanisms. For a more detailed description of the methods applied in this study we refer to the Methods section below.



**Figure 1.** Schematic of the CycleGAN model, showing the two generator-discriminator pairs that learn to translate samples from the ESM simulations to the ERA5 reanalysis (grey) and vice versa (yellow). Training the two generators to learn inverse mappings of each other allows to enforce cycle-consistency in the translation of the unpaired samples, i.e.  $x \rightarrow G(x) \rightarrow F(G(x)) \rightarrow \tilde{x} \approx x$  and vice versa for  $y$ . As described by Zhu et al. (2017), the cycle-consistency loss (Eq. 5) is motivated from natural language translation, where one should arrive at the same sentence after translating it into another language and back. In the training context, this has been found to improve the stability and to prevent typical problems in adversarial networks, such as mode collapse, where every input would be mapped to the same output image (Zhu et al., 2017).

## 2 Results

### 2.1 Correcting temporal distributions

When comparing the spatial precipitation fields from CM2Mc-LPJmL with the ERA5 data, large biases are evident, especially in the tropics, where a pronounced double-peaked Intertropical Convergence Zone of CM2Mc-LPJmL can be seen (Fig. 2a). The more complex and higher-resolution – yet computationally much more expensive – GFDL-ESM4 model exhibits a similar spatial pattern of bias, although with a reduced southern peak (Fig. 2b).

We evaluate our method against quantile mapping, which a state-of-the-art method to correct temporal distributions (Fig. 2c). The GAN shows a strongly improved skill overall, and especially in correcting the double-peaked ITCZ (Fig. 2d), compared to quantile mapping, but also compared to GFDL-ESM4 model.

This is also summarized in the averaged absolute value of the mean error (ME) shown in the spatial plots (Table 1). Here, the GAN shows the strongest error reduction compared to QM and GFDL-ESM4, reducing the error of CM2Mc-LPJmL by 75% for annual and between 72% to 64% for seasonal time series. We include the results of two additional ESMs from CMIP6, the MPI-ESM1-2-HR and the CESM2 model, for comparison with GFDL-ESM4 in the SI (Table S1). The ME of the MPI-ESM1-2-HR model is higher than for GFDL-ESM4 while the CESM2 shows lower bias. The average ME of CESM2, however, remains higher than our GAN-based post-processed CM2Mc-LPJmL model.

In addition to the mean error we also evaluate the difference in the 95th percentile of the precipitation above a threshold of 0.5 [mm/day] per grid cell. The spatial plots are shown in Figs. S5-S9 and summarized as absolute averages in Table S2. Again, the GAN outperforms the other baseline methods for annual and seasonal time series, reducing biases between 59.76 and 49.11%.

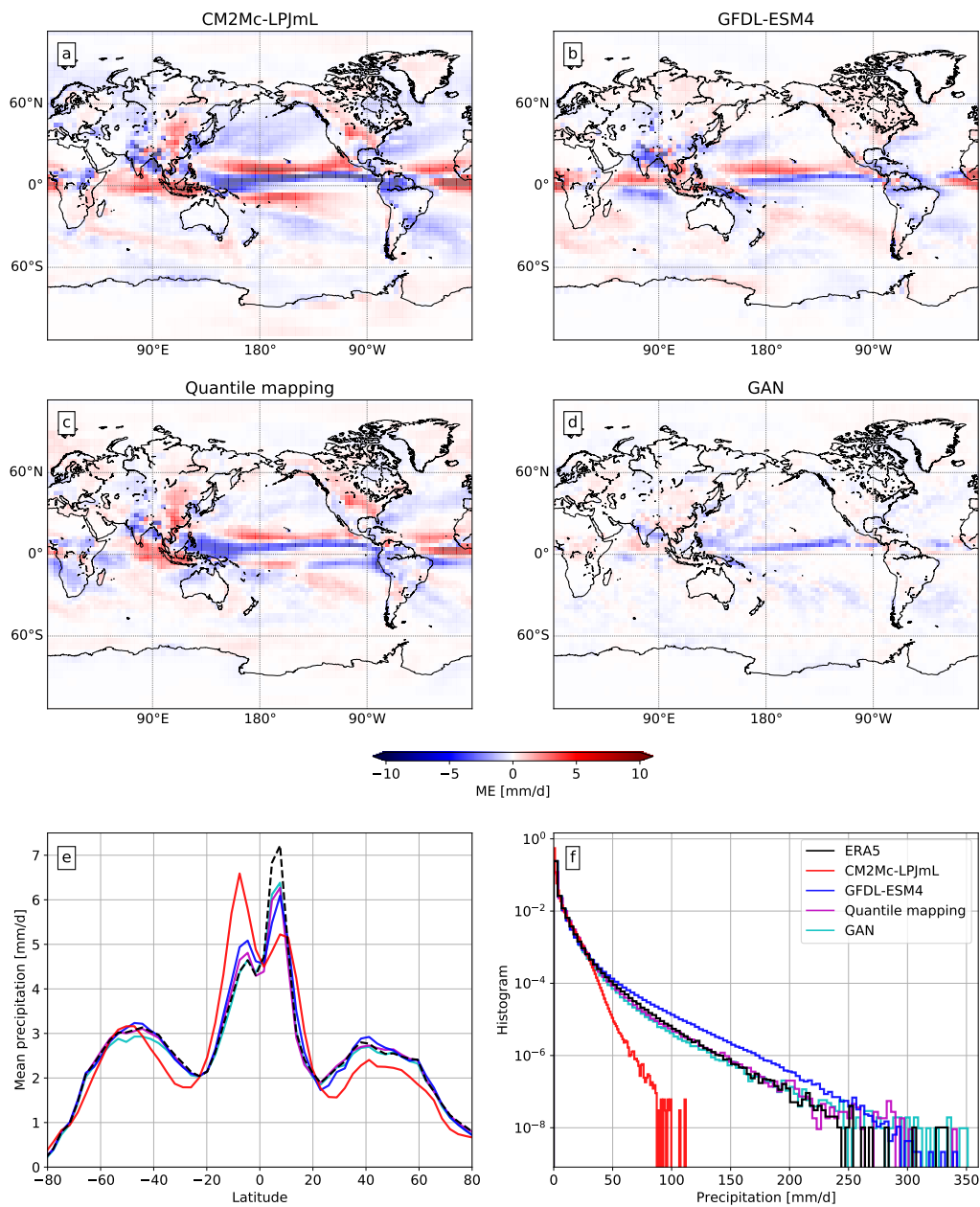
Also from latitude profiles it can be quantitatively inferred that the GAN outperforms quantile mapping especially regarding the correction of the double-peaked ITCZ, and also that the GAN-processed fields is closer to the ERA5 data than the GFDL-ESM4 simulations, especially in the tropics (Fig. 2e).

Regarding the globally averaged temporal distributions, we infer an under-representation of heavy precipitation values in CM2Mc-LPJmL and an over-representation in GFDL-ESM4. QM and our GAN-based method perform similarly well in correcting the distributions over the entire range of precipitation values (Fig. 2f).

**Table 1.** The averaged absolute value of the grid-cell-wise mean error (ME) for the raw CM2Mc-LPJmL and GFDL-ESM4 models, as well as for the QM- and GAN-based post-processing, using the CM2Mc-LPJmL output as input. The bias reduction relative to the raw CM2Mc-LPJmL model is given in percentage. Note that the GAN shows the largest reduction of the absolute ME in all cases, with more than 75% improvement relative to the raw CM2Mc-LPJmL for the annual fields.

Season	CM2Mc-LPJmL	GFDL-ESM4	%	QM	%	GAN	%
Annual	0.769	0.448	41.7	0.218	71.7	<b>0.191</b>	<b>75.2</b>
DJF	0.915	0.544	40.5	0.664	27.4	<b>0.256</b>	<b>72</b>
MAM	0.886	0.603	31.9	0.567	36.4	<b>0.268</b>	<b>69.8</b>
JJA	0.963	0.589	38.8	0.704	26.9	<b>0.270</b>	<b>72</b>
SON	0.823	0.508	38.3	0.552	32.9	<b>0.294</b>	<b>64</b>





**Figure 2.** Comparison of global mean error maps over the JJA season, long-term precipitation statistics based on latitude-profiles and relative frequency histograms. Mean errors of (a) CM2Mc-LPJmL, (b) GFDL-ESM4, (c) QM-based and (d) GAN-based post-processing methods applied to the CM2Mc-LPJmL output. The mean error is computed with respect to the ERA5 reanalysis data. The largest errors are in the tropics, where also the largest mean precipitation values are observed (see panel (e)). The GAN shows the largest error reduction, strongly reducing the double-peaked ITCZ in the tropics. Quantile mapping, on the other hand, is not able to remove the ITCZ bias. See Figs. S1–S4 for corresponding figures for annual time series, as well as the other three seasons. (e) Precipitation rates averaged over time and longitudes and relative frequency histograms (f) are shown for ERA5 data (black), CM2Mc-LPJmL (red), GFDL-ESM4 (blue), quantile mapping (magenta) and the GAN (cyan). The GAN applied to the CM2Mc-LPJmL output corrects the double-peaked ITCZ as well as the histogram over the entire range of precipitation rates.

## 2.2 Correcting spatial patterns

We continue with assessing the ability of our correction method to improve the spatial structure of the ESM precipitation output. Most importantly, we investigate to which degree the characteristic high-frequency spatial variability of precipitation which is not represented well in the CM2Mc-LPJmL model output, can be improved (see Fig. 3 for some example fields). To quantify this spatial intermittency in the precipitation fields, we compute the radially averaged power spectral density (PSD) following (D. Harris et al., 2001; Sinclair & Pegram, 2005; Ravuri et al., 2021). First, the PSD is computed for each daily spatial precipitation field and then the mean is taken over the resulting spectrograms, shown in Fig. 3e. While the CM2Mc-LPJmL precipitation shows a reduced density at high frequencies (i.e., short wavelengths below 1024 km), the GFDL-ESM4 model exhibits an unrealistically high PSD in the same range. Quantile mapping shifts the CM2Mc-LPJmL spectrum towards ERA5, but results in an overshoot in the mid-range and long wavelengths, while the higher frequencies remain underestimated. Only the GAN is able to produce a power spectrum that is consistent with ERA5, especially for short wavelengths, i.e., the high-frequency range that is crucial for precipitation.

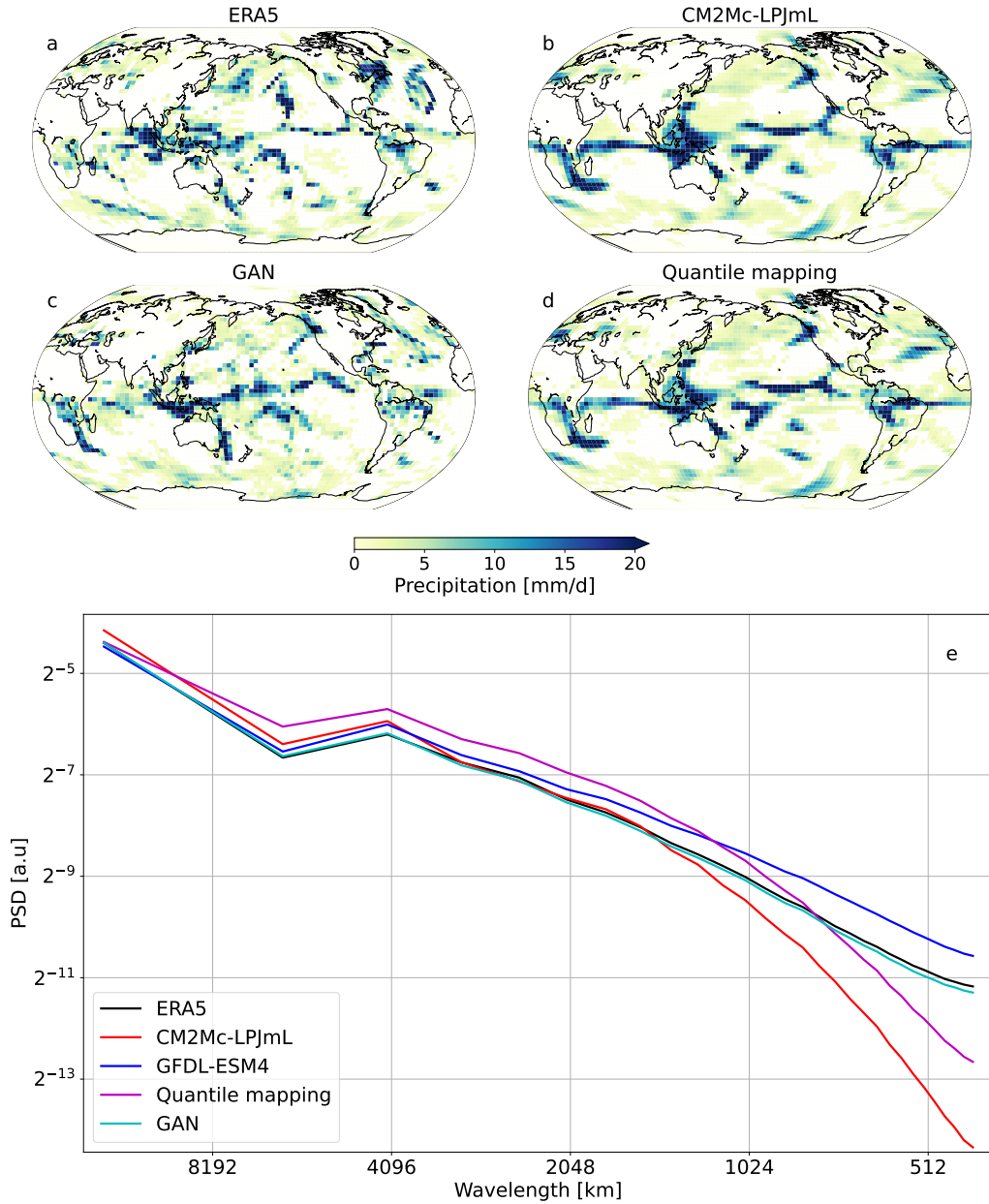
## 2.3 Non-stationary climate scenario

Climate projections under a changing radiative forcing induced by anthropogenic greenhouse gas release constitute an out-of-sample problem: The conditions for which predictions shall be made are different from the conditions for which historical data are available for training. Methods for post-processing or correcting the output of ESMs tasked with such projections hence need to be able to generalize to states that deviate from the historical period, where observations are available. Here, we test our GAN approach for the CMIP6 SSP5-8.5 scenario until the end of the 21st century. The SSP5-8.5 “business as usual” scenario represents an extreme climate scenario in CMIP6, with the strongest increase in CO<sub>2</sub>. This scenario has been chosen to test how well the GAN model can capture the non-stationarity in this extreme case.

The CM2Mc-LPJmL and GFDL-ESM4 models both show monotonically increasing global mean precipitation with similar trends over the current century (Fig. 4a), which is in agreement with other studies (IPCC, 2021). In contrast, the unconstrained GAN, trained on the historical period, does – as expected – not exhibit an increase in average global precipitation, since it is by itself not able to generalize to the changing boundary conditions given by higher greenhouse gas concentrations and temperatures.

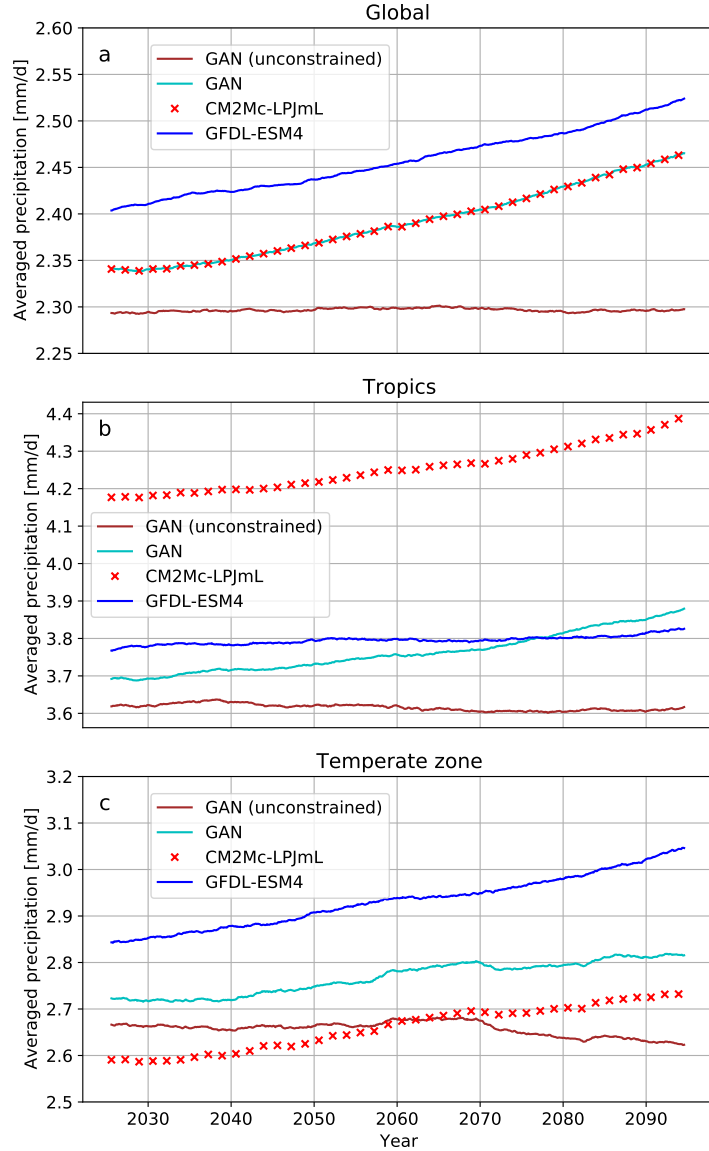
In the tropics (23° S to 23° N), GFDL-ESM4 remains overall lower in mean precipitation than CM2Mc-LPJmL, while also exhibiting a much less pronounced increase over the entire period (Fig. 4b). For the temperate zones from 40° N/S to 60° N/S, the GFDL-ESM4 model shows an overall higher mean precipitation with a slightly stronger positive trend than CM2Mc-LPJmL (Fig. 4c).

By construction of the constraint introduced in Eq. 8, the GAN-processed precipitation is identical to the increasing global average of the CM2Mc-LPJmL output (Fig. 4a). Without the constraining layer added to the GAN, however, the GAN-processed precipitation stays relatively constant without a substantial trend. In both tropical and temperate zones, the constrained GAN corrects the precipitation towards the more complex and higher-resolution GFDL-ESM4, while following the trend of the CM2Mc-LPJmL model. Again, the unconstrained model remains relatively constant in both cases, with a small decrease over time in the temperate zone. Note that the GFDL-ESM4 does not represent a ground truth, but only one realisation of a possible Earth system trajectory, for comparison. This can be seen by the differing trends of two other CMIP6 models in Fig. S13. It should, however, be expected that the precipitation output from the CMIP6 models is much more realistic than the raw precipitation from the comparably low-resolution CM2Mc-LPJmL model. The CMIP6 model GFDL-ESM4 also appears to be calibrated well with respect to large-scale



**Figure 3.** Qualitative and quantitative comparison of the intermittency in daily precipitation above 1 mm/day, on the same date (25th December 2014), for the (a) ERA5 reanalysis, (b) CM2Mc-LPJmL model, (c) GAN-based and (d) QM-based post-processing. The CM2Mc-LPJmL precipitation field (b) corresponds to an input of the GAN-generator which transforms it into the field shown in panel (c). The discriminator network then classifies whether the GAN output (c) or the ERA5 field (a) was generated artificially. Visually, the GAN substantially improves the spatial intermittency seen in ERA5, whereas applying QM does not lead to improved intermittency. Note that the modelled fields are not expected to be point-wise similar to the ERA5 ‘ground truth’ (a), since these are time slices from climate projection runs. (e) The spatial power spectral density (PSD) of the different precipitation fields, averaged radially in space and over time. For ERA5 reanalysis (black), CM2Mc-LPJmL (red), GFDL-ESM4 (blue), quantile mapping (magenta) and the GAN (cyan). Note that only GAN-based post-processing of the CM2Mc-LPJmL model yields an accurate PSD across all spatial scales.

averages over the historical test period, as can be seen in Fig. S12, in which the GAN shows improvements over the CM2Mc-LPJmL inputs.



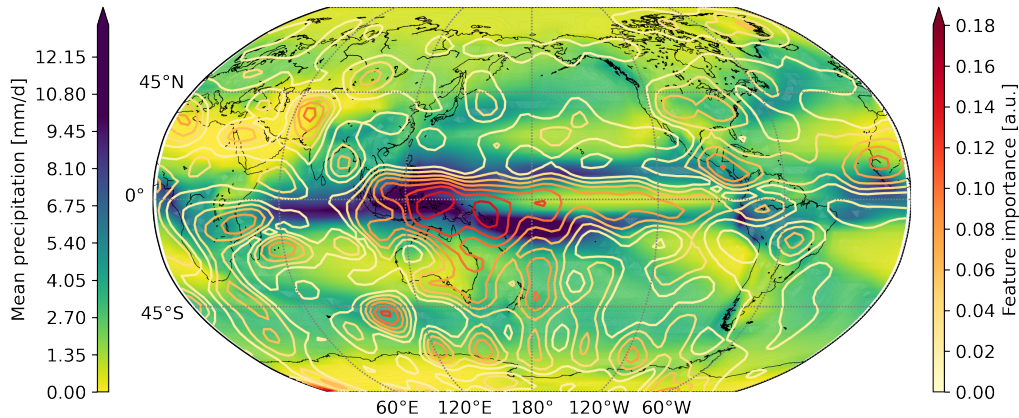
**Figure 4.** Large-scale trends as a three year rolling-mean of monthly and spatially average precipitation for the CMIP6 SSP5-8.5 scenario. For (a) global data, (b) the tropics and (c) temperate zone, of the CM2Mc-LPJmL (red crosses) and GFDL-ESM4 (blue) models, as well as the constrained (cyan) and unconstrained (brown) GANs. Only by adding the physical constrained to preserve the global precipitation amount per timestep enables the GAN (cyan) to follow the transient dynamics of the non-stationary climate scenario.

## 2.4 Interpretability of the GAN-based correction

We investigate in the following whether the GAN has learned an ESM output correction that is also physically reasonable. The attribution maps are computed with SmoothGrad

for each prediction of the discriminator  $D_Y$ , with daily CM2Mc-LPJmL precipitation fields given as input. The discriminator has been trained to distinguish between reanalysis (ERA5) and GAN-processed precipitation fields and we are interested to see which spatial regions in the ESM output the discriminator regards as most important for the distinction. These regions then need to be corrected the most by the generator, implying where the most pronounced biases of CM2Mc-LPJmL are.

The temporal average of the CM2Mc-LPJmL precipitation is shown in Fig. 5 together with the absolute value of the attribution map as contour lines. The regions of highest importance are shown in red and coincide with the region in the western Pacific where the strongest biases and in particular the double-peaked ITCZ of CM2Mc-LPJmL are located (as shown in Fig. 2 and Fig. S1). Although the GAN is trained on daily precipitation fields, it has thus learned to identify regions that show biases occurring on interseasonal to interannual scales.



**Figure 5.** Annual average of daily precipitation fields from CM2Mc-LPJmL (color shading with scale according to the colorbar on the left) together with attribution maps (contour lines with color scale according to colorbar on the right). Note that we applied a Gaussian filter to the attribution maps to further reduce the noise. A standard deviation  $\sigma = 1.5$  for the filter was found to give robust results. The pacific region in the tropics shows the highest annual mean precipitation, and also the highest feature importance. The same region also exhibits the largest bias of CM2Mc-LPJmL, see in Fig. 2. Note that especially the double-ITCZ bias is a common and long-standing problem in the precipitation output of many general circulation models (Tian & Dong, 2020).

### 3 Discussion

We have introduced a physically constrained generative adversarial network that, combined with the computationally lightweight and efficient CM2Mc-LPJmL Earth system model, is able to produce highly realistic precipitation simulations at low computational costs.

Our method improves the ESM output in two ways: (i) the temporal distributions of the CM2Mc-LPJmL model precipitation, as well as (ii) the spatial patterns and in particular the spatial intermittency of the CM2Mc-LPJmL model precipitation. Our approach is evaluated against quantile mapping (Cannon et al., 2015) and the much more advanced CMIP6 GFDL-ESM4 model, (Krasting et al., 2018) taking ERA5 reanalysis data as ground

truth. Note that any other, and especially purely observational, precipitation dataset with sufficient temporal resolution could readily be used instead.

Given that the training samples are unpaired as a result of the chaotic nature of observed and simulated Earth system trajectories, a comparison of single prediction-target pairs is not possible. We therefore evaluate the GAN performance on long-term summary statistics over the entire test set period. When evaluating the skill to improve temporal distributions, we find that our proposed method outperforms both baselines, showing the lowest mean errors and the smallest difference in the 95th precipitation percentile. The improvement over quantile mapping is especially pronounced for seasonal time series, where only our method successfully removes the double-peaked ITCZ of the CM2Mc-LPJmL model. This is in contrast to the results by (François et al., 2021), who report a comparable skill of their CycleGAN implementation with quantile mapping for regional climate simulations. Our method corrects relative frequency histograms over the entire range of precipitation values, similarly well to QM, which is designed for this task.

Crucially, our GAN-based approach also improves the spatial structure of the ESM precipitation fields, which is not possible with traditional approaches. The GAN yields realistically intermittent spatial patterns that are characteristic for precipitation on all resolved scales, and in this regard outperforms both the quantile-mapping-based post-processing and the comprehensive, high-resolution GFDL-ESM4 model. These results show that our method, combined with the computationally lightweight and efficient CM2Mc-LPJmL ESM, can produce precipitation fields that are at least comparable to state-of-the-art, and much more computationally expensive CMIP6 models.

We applied our method to the strongly non-stationary SSP5-8.5 CMIP6 climate scenario until 2100 to test the GAN’s ability to capture these non-stationarity and the transient dynamics. The unconstrained GAN trained on observations does not generalize to the unseen climate state. It does not show an increase in global mean precipitation, as one would expect from the thermodynamic Clausius-Clapeyron relation and as seen in the numerical ESMs (Allan & Soden, 2008; Donat et al., 2013; Guerreiro et al., 2018; Traxl et al., 2021). This can be explained by the fact that the precipitation of the future scenario lies well outside the training distribution. To solve this and help the GAN to generalize to this kind of out-of-sample prediction, a physical constraint to preserve the global precipitation amount of the ESM in each time step was introduced as additional network layer in the GAN. The global constraint allows the GAN to improve the precipitation regionally by accounting for local characteristics, while producing the same global mean as the ESM by construction. Conserving a physical quantity that is simulated numerically, such as the global precipitation sum in our study, also means that it cannot be improved with respect to observations by definition of the constraint. The global precipitation trend can, however, be expected to be represented comparably well in the numerical ESM through thermodynamic processes. Adding this constraint enables the GAN to follow the non-stationary, transient dynamics of the SSP5-8.5 scenario.

The generator architecture in this study is deterministic, producing the same input-output-pairs once it is trained. This enables run-to-run reproducibility, where uncertainties of the ESM can then be quantified through ensemble runs. Since the training itself is stochastic, one can create an ensemble to estimate the uncertainties resulting from GAN training (see Fig. S14). A potential direction for future research could be to develop a stochastic model that directly learns the uncertainties.

We demonstrate how feature attribution from interpretable Artificial Intelligence can be applied for a GAN, enabling a physical interpretation of this deep learning model. We find that the discriminator part of the GAN has learned to identify those regions for its decisions that are critical also from a physical perspective. These regions highlighted by our GAN interpretation are the ones with the highest absolute errors of the raw CM2Mc-LPJmL, and are known to be the most problematic for ESM precipitation in general. Namely, the

tropical Pacific Ocean was found to be of highest importance for the discriminator. In this region, the particularly heavy precipitation is often caused by deep convection-driven clouds, which are difficult to model numerically (Tian & Dong, 2020). The sensitivity of the discriminator in the Pacific region also explains the effectiveness of our generator network to reduce the double-peaked ITCZ bias. This is the region where the generator needs to modify the CM2Mc-LPJmL precipitation field most in order to avoid rejection by the discriminator. The results indicate that the GAN has successfully learned the long-term statistics while being trained on samples of much shorter time scales. This makes GANs particularly suitable for climate applications, where training samples and the statistics of interest are often on very different time scales.

The main contribution of our approach is the efficient simulations of highly realistic precipitation fields, by combining a physically constrained GAN with an ESM of reduced complexity. Producing similarly realistic fields purely numerically would require much more computational resources. For comparison, our post-processed CM2Mc-LPJmL ESM takes about 0.5 hours to compute a model year using 28 CPUs, whereas the much more complex GFDL-ESM4 requires 2 hours computational time on 1000 CPUs for a model year (Krasting et al., 2018). This corresponds to an increased computational efficiency by roughly two orders of magnitude, keeping in mind that GFDL-ESM4 produces higher resolution output. The time the GAN post-processing takes is negligible in comparison, taking 0.35 seconds per model year on a V100 GPU and 37.17 seconds on a single CPU. The quantile mapping is similarly efficient taking 0.59 seconds per model year on a CPU.

Based on our findings, there are several directions for extending our method. Down-scaling applications that increase the resolution of the ESM could be a direction for future research. Conditioning the generator by adding variables that are physically linked to precipitation, such as humidity, temperature, or wind, could further improve our method. The precipitation data, improved by our method, may be used as input to other stand-alone Earth system components such as vegetation, that require realistic climate input.

## Acknowledgments

The authors would like to thank the referees for their helpful comments and suggestions. NB and PH acknowledge funding by the Volkswagen Foundation, as well as the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research and the Land Brandenburg for supporting this project by providing resources on the high performance computer system at the Potsdam Institute for Climate Impact Research. MD acknowledges funding by the Volkswagen Foundation project POEM-PBSim. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting FS. NB acknowledges further funding by the Federal Ministry of Education and Research under grant No. 01LS2001A.

## Data availability

The ERA5 reanalysis data is available for download at the Copernicus Climate Change Service (C3S) (<https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview> and <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-preliminary-back-extension?tab=overview>). Output data from the CM2Mc-LPJmL model is available at <https://doi.org/10.5281/zenodo.4683086> (Drüke, 2021). The CMIP6 data can be downloaded at <https://esgf-node.llnl.gov/projects/cmip6/>.

## Code availability

For the CM2Mc-LPJmL model code see <https://doi.org/10.5281/zenodo.4700270> (Drüke, Petri, et al., 2021). The Python code for processing and analysing the data, together with the PyTorch Lightning (Falcon et al., 2019a, 2019b) code for training is available as a compute capsule at Code Ocean: <https://doi.org/10.24433/CO.2750913.v1> (Hess et al., 2022).

## Competing interests

The authors declare no competing interests.

## Authors contribution

PH and NB conceived the research and designed the study with input from all authors. PH performed the numerical analysis. MD conducted the CM2Mc-LPJmL experiments. All authors interpreted and discussed the results. PH wrote the manuscript with input from all authors.

## Materials and Methods

### The Earth system model CM2Mc-LPJmL

The coupled Earth system model CM2Mc-LPJmL v1.0 (Drüke, von Bloh, et al., 2021) combines the coarse-grained but relatively fast atmosphere and ocean model CM2Mc (Galbraith et al., 2011) with the state-of-the-art dynamic global vegetation model (DGVM) LPJmL5 (Schaphoff et al., 2018a, 2018b; Von Bloh et al., 2018).

CM2Mc is a coarser ( $3^\circ \times 3.75^\circ$  latitude-longitude) configuration of the Climate Model CM2 (Milly & Shmakin, 2002), which has been developed at the Geophysical Fluid Dynamics Laboratory (GFDL). The original configuration of CM2Mc includes the Modular Ocean Model 5 (MOM5) and the global atmosphere and land models AM2-LM2 or AM2-LM (Anderson et al., 2004) with static vegetation. In CM2Mc-LPJmL, the land component LM/LM2 is replaced by the dynamic global vegetation model LPJmL5, while AM2 and MOM5 remain dynamically coupled to the model framework. The Flexible Modeling System (FMS) developed by GFDL connects all different model compartments and calculates the fluxes between them.

The state-of-the-art and thoroughly validated DGVM LPJmL (Lund-Potsdam-Jena managed Land) simulates global surface energy balance, water fluxes and carbon stocks and fluxes for natural and managed land. Being forced by climate and soil data, LPJmL simulates the impact of bioclimatic limits and effects of heat, productivity and fire on plant mortality to determine the establishment, growth, competition and mortality for different plant functional types (PFTs) in natural vegetation and crop functional types (CFTs) on managed land. Since its original implementation (Sitch et al., 2003) the model now incorporates a water balance (Gerten et al., 2004), agriculture (Bondeau et al., 2007), wildfire in natural vegetation (Thonicke et al., 2010; Drüke et al., 2019), and the impact of multiple climate drivers on phenology (Forkel et al., 2014, 2019).

In CM2Mc-LPJmL, the fluxes simulated by LPJmL depend, of course, on the precipitation modelled by AM2. As a stand-alone model LPJmL has been mainly calibrated with respect to reanalysis, and a similarly accurate precipitation output within CM2Mc-LPJmL would hence be favorable to maintain consistency and to obtain realistic surface fluxes from LPJmL. For the overall performance of CM2Mc-LPJmL, realistically simulated precipitation fields are therefore crucial. This motivates the work presented below, where we use a specific kind of GAN to transform the AM2 precipitation fields toward fields that are indistinguishable from ERA5 precipitation fields (see below).



The model experiments of this paper are consistent with (Drüke, von Bloh, et al., 2021). After a 5000-year stand-alone LPJmL spin-up, a second fully coupled spin-up under pre-industrial conditions without land use was performed for 1250 model years. In this way we ensure that the model starts from a consistent equilibrium between the long-term soil carbon pool, vegetation, ocean, and climate.

The subsequent transient historic phase of the model is performed from 1700-2018, using historic land use data from 1700 (Fader et al., 2010) and historic concentrations of greenhouse gases, solar radiation, ozone concentrations and aerosols from 1860, which were kept at pre-industrial conditions beforehand.

From 2019 until 2100 the model is forced by constant land use from the year 2018 and CO<sub>2</sub> equivalents of the atmospheric forcing prescribed in the CMIP6 SSP5-8.5 (“business as usual”) climate scenario that assumes a continued increase in CO<sub>2</sub> emissions.

### Cycle-consistent generative adversarial networks

Generative adversarial networks (GANs) are designed to learn a target distribution  $p_y(y)$  through a two-player “minimax” game between a generator  $G$  and a discriminator  $D$  (Goodfellow et al., 2014). The generator network is trained to transform an input  $x \in X$  to values that approximate samples from a target domain  $y \in Y$ , i.e. the generator is trained to learn the mapping  $G : X \rightarrow Y$ . Samples from the generator and the target dataset are then shown to the discriminator, which classifies their origin. In this way, the generator and discriminator compete against each other, thereby improving the quality of the generated samples. The training can be formulated as

$$G^* = \min_G \max_D \mathcal{L}_{GAN}(D, G), \quad (1)$$

where  $G^*$  is the optimal generator and  $\mathcal{L}_{GAN}(D, G)$  is the loss function defined as

$$\mathcal{L}_{GAN}(D, G) = \mathbb{E}_{y \sim p_y(y)}[\log(D(y))] + \mathbb{E}_{x \sim p_x(x)}[\log(1 - D(G(x)))]. \quad (2)$$

In our situation,  $X$  and  $Y$  correspond to the sets containing precipitation fields from the CM2Mc-LPJmL Earth system model and ERA5 reanalysis, respectively (samples are shown in Fig. 3). In the above formulation, GANs have often been found to suffer from instabilities and difficulties to generalize to distributions of higher dimensionality, such as in image-to-image translation without pairwise matching samples. One reason for the instabilities is the highly under-constrained mapping to be learned by the generator. To alleviate this problem, cycle-consistent GANs have been proposed recently (Zhu et al., 2017). They aim to constrain the space of mappings by training a second pair of generator and discriminator networks, which learns the inverse mapping  $F : Y \rightarrow X$ . A schematic of the cycle-consistent GAN model is shown in Fig. 1. Both generators should perform bijective (i.e., one-to-one) mappings (Zhu et al., 2017) and are therefore trained at the same time, together with a regularization term that enforces consistency of translation cycles, i.e.  $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$  and vice versa for  $y$ . The corresponding loss functions are then

$$\mathcal{L}_{X \rightarrow Y}(G, D_Y) = \mathbb{E}_{y \sim p_y(y)}[\log(D_Y(y))] + \mathbb{E}_{x \sim p_x(x)}[\log(1 - D_Y(G(x)))], \quad (3)$$

and similarly,

$$\mathcal{L}_{Y \rightarrow X}(F, D_X) = \mathbb{E}_{x \sim p_x(x)}[\log(D_X(x))] + \mathbb{E}_{y \sim p_y(y)}[\log(1 - D_X(F(y)))]. \quad (4)$$

The cycle-consistency loss is given by

$$\begin{aligned} \mathcal{L}_{cycle}(G, F) = & \mathbb{E}_{x \sim p_x(x)} [\|F(G(x)) - x\|_1] \\ & + \mathbb{E}_{y \sim p_y(y)} [\|G(F(y)) - y\|_1]. \end{aligned} \quad (5)$$

The full loss function then reads

$$\begin{aligned} \mathcal{L}(G, F, D_X, D_Y) = & \mathcal{L}_{X \rightarrow Y}(G, D_Y) \\ & + \mathcal{L}_{Y \rightarrow X}(F, D_X) \\ & + \lambda \mathcal{L}_{cycle}(G, F), \end{aligned} \quad (6)$$

which is solved as

$$G^*, F^* = \min_{G, F} \max_{D_X, D_Y} \mathcal{L}(G, F, D_X, D_Y). \quad (7)$$

We adopt the architecture from Zhu et al. (2017) and optimize the networks with Adam (Kingma & Ba, 2014), using a learning rate of  $2e^{-4}$  for both the generator and the discriminator networks and set  $\lambda = 10$ . Following Zhu et al. (2017) we set the batch size to 1 and train the models for 250 epochs, logging the 50 best performing generators every 10 epochs. The training takes about 5.25 days on a NVIDIA V100 GPU with 32 GB memory. After training the final generator is determined by evaluation on the test set.

### Neural network architectures

The generator architecture is based on a variant of convolutional residual networks (He et al., 2016). Convolutional neural networks (CNNs) are commonly employed to process image data. CNNs transform the input data through stacked layers of trainable convolutional filters that are followed by a non-linear activation functions thereby learning to extract spatial patterns. For a more detailed introduction see, e.g., (Goodfellow et al., 2016). Adopting the naming convention from (Johnson et al., 2016; Zhu et al., 2017). `c7s1-k` denotes a layer with a  $7 \times 7$  convolution followed by instance normalization and ReLU activation with  $k$  filters, a stride 1 and reflection padding. `dk` represents a layer with  $3 \times 3$  convolutions, instance normalization, ReLU activation,  $k$  filters and stride 2. `Rk` are residual blocks with a  $3 \times 3$  convolutional layer and  $k$  filters. `uk` denotes a layer with  $3 \times 3$  fractional-strided convolutions, instance normalization, ReLU activation,  $k$  filters and stride 1/2. The generator architecture with 6 residual blocks is then

$$x_{\text{in}} \rightarrow \text{c7s1-64} \rightarrow \text{d128} \rightarrow \text{d256} \rightarrow \underbrace{[\text{R256} \rightarrow]}_{\times 6} \text{u128} \rightarrow \text{u64} \rightarrow \text{c7s1-3} \rightarrow y_{\text{out}},$$

where  $x_{\text{in}}$  is the input of the generator and  $y_{\text{out}}$  the output. The discriminator architecture is based on the PatchGAN (Isola et al., 2017). Denoting a  $4 \times 4$  convolutional layer with  $k$  filters, instance normalization (except for the first layer), leaky ReLU with slope 0.2 and a stride of 2 with `Ck`. The full architecture of the discriminator is

$$x_{\text{in}} \rightarrow \text{C64} \rightarrow \text{C128} \rightarrow \text{C256} \rightarrow \text{C512} \rightarrow y_{\text{out}}.$$

### Generator constraint

To enable a better generalization of the GAN to climate states not seen during training, and hence in particular to address the out-of-sample problem imposed by the changing

radiative forcing due to anthropogenic greenhouse gas emissions, we introduce the physical constraint of preserving the total global precipitation amount of the CM2Mc-LPJmL model input. That is, we add an additional layer to the generator network after training, which re-scales each output  $y_i$  at each grid point  $i$  as

$$\tilde{y}_i = y_i \frac{\sum_i^{N_{\text{grid}}} x_i}{\sum_i^{N_{\text{grid}}} y_i}, \quad (8)$$

where  $N_{\text{grid}}$  is the total number of grid-points,  $x_i$  the CM2Mc-LPJmL precipitation input and  $\tilde{y}_i$  the constrained output. The motivation of the constraint is that it gives the GAN freedom to change the precipitation locally and to redistribute it in space, while forcing it to follow the global trend prescribed by the ESM. The global trend has been found to be well represented in the ESM, where noise and biases found on small time and spatial scales are averaged out (Drüke, von Bloh, et al., 2021). Also in observations, it has recently been shown that the physically based Clausius-Clapeyron relation, suggesting a 7% increase in precipitation per degree of warming, holds very well in terms of global averages, despite pronounced regional deviations (Traxl et al., 2021).

## Training

We use daily precipitation from the European Center for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) product (Hersbach et al., 2020) as a training target and ground truth for evaluation. This reanalysis is produced by the Copernicus Climate Change Service (C3S) at ECMWF, combining a large range of satellite- and land-based observations with high-resolution simulations through state-of-the-art data assimilation techniques (Courtier et al., 1994; Hersbach et al., 2020). The original resolution is 30km horizontally in space and hourly in time, spanning the period from 1950 to present. For this study the data is aggregated to daily precipitation sums and re-gridded, following (Rasp et al., 2020; Beck et al., 2019), by bilinear interpolation using the xESMF package (Zhuang et al., 2020), in order to match the resolution of CM2Mc-LPJmL. We split the ESM and ERA5 datasets into the training period 1950-2000 and the test period 2001-2014 (for which also the GFDL-ESM4 data is available), with 18615 and 5110 daily samples, respectively. Model simulations from 2019-2100 are used to test the generalization of the network with a CO<sub>2</sub> forcing according the CMIP6 SSP5-8.5 (“business as usual”) climate scenario, which assumes a continued increase in CO<sub>2</sub> emissions. Following Zhu et al. (2017), we replace the log likelihood by a least-squares loss, which has been found to improve the training. The GAN loss in Eq. 2 is then minimized by both  $G$  and  $D$ , with a loss  $\mathbb{E}_{x \sim p_x(x)}[(D(G(x)) - 1)^2]$  for  $G$  and  $\mathbb{E}_{y \sim p_y(y)}[(D(y) - 1)^2] + \mathbb{E}_{x \sim p_x(x)}[(D(G(x)))^2]$  for the discriminator  $D$ . We apply a log-transform to the input data with  $\tilde{x} = \log(x + \epsilon) - \log(\epsilon)$  following (Rasp & Thuerey, 2021), where  $\tilde{x}$  is the transformed precipitation and  $\epsilon = 0.0001$ . We further normalize the data to the interval  $[-1, 1]$ , which was found to improve the training performance. Once trained, the generator takes only about ten seconds on a NVIDIA V100 GPU to process the test set ESM precipitation.

## Baselines

We compare our method to quantile mapping, implemented with the xClim package (Logan et al., 2021), and also carry out comparisons to the raw output of the more advanced CMIP6 climate model GFDL-ESM4 (Krasting et al., 2018). The latter uses AM4 (Zhao et al., 2018a, 2018b), a more recent and substantially more complex version of the atmosphere model AM2 used in CM2Mc-LPJmL (GFDL Global Atmospheric Model Development Team et al., 2004), with a substantially higher spatial resolution and strongly improved parameterizations of subgrid-scale processes. These improvements of course come at the expense of substantially increased computational costs. The motivation here is to see whether a comparably simple atmospheric general circulation model (GCM) such as AM2 can be com-

bined with the proposed GAN model in order to yield similar results as a comprehensive state-of-the-art atmospheric GCM such as AM4, at a fraction of the computational costs. Quantile mapping uses the empirical cumulative distribution functions of simulated and observed precipitation to transform the simulated values into the corresponding quantiles derived from observations. Before computing the cumulative distribution function, following (Cannon et al., 2015), we detrend the historical time series, assuming a linear trend. As an error metric to compare our methods we apply the mean error (ME), which is defined as

$$\text{ME} = \frac{1}{N} \sum_{t=1}^{N_{time}} (x_t - y_t) = \frac{1}{N} \sum_{t=1}^{N_{time}} x_t - \frac{1}{N} \sum_{t=1}^{N_{time}} y_t, \quad (9)$$

where  $x_t$  and  $y_t$  are the simulated and observed precipitation at time  $t$  for a given grid cell and  $N_{time}$  the number of time steps in the test set. Note that the ME is used to evaluate the differences in the time averages per grid cell, as can be seen on the right-hand side of Eq. 9.

### Model transparency

Neural network models are often regarded as black boxes. Since it is important for many applications to be able to explain the neural network’s prediction, the emergent fields of interpretable (Murdoch et al., 2019; Toms et al., 2020) and explainable Artificial Intelligence (Sundararajan et al., 2017; Montavon et al., 2019) aim to improve the transparency.

Many methods for interpreting neural networks are specifically designed for classification problems (Goodfellow et al., 2016). In the GAN framework, the discriminator network performs such a classification task in distinguishing between generated and real images. Hence, suitable interpretability methods can be applied, even though entire GAN is build for the much more complex generative task. Being able to interpret the GAN increases the transparency and trust, since it ensures that the model has learned to identify physically reasonable input features. To our knowledge, we are the first to apply an interpretability method in such a way, i.e., to test the physical consistency of the GAN training.

Here, we use the gradient-based method SmoothGrad (Smilkov et al., 2017) to interpret the discriminator network  $D_Y$  that has learned to classify ERA5 and generated precipitation fields. An attribution map  $\phi$  is computed by taking the gradient of the neural network  $D_Y$  with respect to its input  $y$ ,

$$\phi(D_Y, y) = \frac{\partial D_Y(y)}{\partial y}, \quad (10)$$

showing for each input grid cell how much the prediction will change with respect to its input, i.e. how sensitive it is to perturbations of the input. It has been observed that using only the gradient of the input, however, tends to give rather noisy attribution maps. Therefore, Smilkov et al. (2017) proposed a technique to reduce the noise, by adding it to the network’s input and averaging the gradient over a sample size, e.g. here  $N = 10$ , as

$$\hat{\phi}(D_Y, y) = \frac{1}{N} \sum_{i=1}^N \phi(y + \epsilon_i), \quad (11)$$

where the noise is sampled from a Gaussian distribution  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ .

## References

- Allan, R. P., & Soden, B. J. (2008). Atmospheric warming and the amplification of precipitation extremes. *Science*, *321* (5895), 1481–1484.
- Anderson, J. L., Balaji, V., Broccoli, A. J., Cooke, W. F., Delworth, T. L., Dixon, K. W., . . . Wyman, B. L. (2004). The new GFDL global atmosphere and land model AM2-LM2:

- Evaluation with prescribed SST simulations. *Journal of Climate*, 17(24), 4641–4673. doi: 10.1175/JCLI-3223.1
- Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I., ... Adler, R. F. (2019). MSWEP V2 global 3-hourly 0.1° precipitation: methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, 100(3), 473–500. doi: 10.1175/BAMS-D-17-0138.1
- Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P. (2021). Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, 126(9), 098302.
- Bondeau, A., Smith, P. C., Zaehle, S., Schaphoff, S., Lucht, W., Cramer, W., ... Smith, B. (2007). Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Global Change Biol.*, 13(3), 679–706. doi: 10.1111/j.1365-2486.2006.01305.x
- Boyle, J., & Klein, S. A. (2010). Impact of horizontal resolution on climate model forecasts of tropical precipitation and diabatic heating for the TWP-ICE period. *Journal of Geophysical Research: Atmospheres*, 115(D23). doi: 10.1029/2010JD014262
- Cannon, A. J., Sobie, S. R., & Murdock, T. Q. (2015). Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *Journal of Climate*, 28(17), 6938–6959. doi: 10.1175/JCLI-D-14-00754.1
- Courtier, P., Thépaut, J.-N., & Hollingsworth, A. (1994). A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120(519), 1367–1387.
- Déqué, M. (2007). Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global and Planetary Change*, 57(1-2), 16–26.
- Donat, M. G., Alexander, L. V., Yang, H., Durre, I., Vose, R., Dunn, R. J. H., ... Kitching, S. (2013). Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset. *Journal of Geophysical Research: Atmospheres*, 118(5), 2098–2118. doi: 10.1002/jgrd.50150
- Drüke, M. (2021, April). *Output data for the GMD publication gmd-2020-436*. [Data set] Zenodo. doi: 10.5281/zenodo.4683086
- Drüke, M., Forkel, M., von Bloh, W., Sakschewski, B., Cardoso, M., Bustamante, M., ... Thonicke, K. (2019). Improving the LPJmL4-SPITFIRE vegetation-fire model for South America using satellite data. *Geoscientific Model Development*, 12(12), 5029–5054. doi: 10.5194/gmd-12-5029-2019
- Drüke, M., Petri, S., von Bloh, W., & Schaphoff, S. (2021, April). *Model code for the GMD publication gmd-2020-436 (Version 1.0)*. [Data set] Zenodo. doi: 10.5281/zenodo.4700270
- Drüke, M., von Bloh, W., Petri, S., Sakschewski, B., Schaphoff, S., Forkel, M., ... Thonicke, K. (2021). CM2Mc-LPJmL v1.0: Biophysical coupling of a process-based dynamic vegetation model with managed land to a general circulation model. *Geoscientific Model Development*, 14(6), 4117–4141. doi: 10.5194/gmd-14-4117-2021
- Fader, M., Rost, S., Mueller, C., Bondeau, A., & Gerten, D. (2010, 4). Virtual water content of temperate cereals and maize: Present and potential future patterns. *J. Hydrol.*, 384(3), 218–231. doi: 10.1016/j.jhydrol.2009.12.011
- Falcon, W., et al. (2019a). Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 3, 6.
- Falcon, W., et al. (2019b). *PyTorch Lightning*. GitHub repository. Retrieved from <https://github.com/PyTorchLightning/pytorch-lightning>
- Forkel, M., Carvalhais, N., Schaphoff, S., von Bloh, W., Migliavacca, M., Thurner, M., & Thonicke, K. (2014). Identifying environmental controls on vegetation greenness phenology through model-data integration. *Biogeosciences*, 11(23), 7025–7050. doi: 10.5194/bg-11-7025-2014
- Forkel, M., Drüke, M., Thurner, M., Dorigo, W., Schaphoff, S., Thonicke, K., ... Carvalhais, N. (2019). Constraining modelled global vegetation dynamics and carbon turnover using multiple satellite observations. *Scientific Reports*, 9(1). doi: 10.1038/s41598

- François, B., Thao, S., & Vrac, M. (2021). Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks. *Climate Dynamics*, 57(11), 3323–3353.
- Gagne, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz’96 model. *Journal of Advances in Modeling Earth Systems*, 12(3), e2019MS001896.
- Galbraith, E. D., Kwon, E. Y., Gnanadesikan, A., Rodgers, K. B., Griffies, S. M., Bianchi, D., ... Held, I. M. (2011). Climate variability and radiocarbon in the CM2Mc earth system model. *Journal of Climate*, 24(16), 4230–4254. doi: 10.1175/2011JCLI3919.1
- Gerten, D., Schaphoff, S., Haberlandt, U., Lucht, W., & Sitch, S. (2004, 1). Terrestrial vegetation and water balance - hydrological evaluation of a dynamic global vegetation model. *J. Hydrol.*, 286(1), 249–270. doi: 10.1016/j.jhydrol.2003.09.029
- GFDL Global Atmospheric Model Development Team, Anderson, J. L., Balaji, V., Broccoli, A. J., Cooke, W. F., Delworth, T. L., ... Wyman, B. (2004). The new GFDL global atmosphere and land model AM2–LM2: Evaluation with prescribed SST simulations. *Journal of Climate*, 17(24), 4641–4673. doi: 10.1175/JCLI-3223.1
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2021). Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200092. doi: 10.1098/rsta.2020.0092
- Gudmundsson, L., Bremnes, J. B., Haugen, J. E., & Engen-Skaugen, T. (2012). Downscaling RCM precipitation to the station scale using statistical transformations—a comparison of methods. *Hydrology and Earth System Sciences*, 16(9), 3383–3390.
- Guerreiro, S. B., Fowler, H. J., Barbero, R., Westra, S., Lenderink, G., Blenkinsop, S., ... Li, X.-F. (2018). Detection of continental-scale intensification of hourly rainfall extremes. *Nature Climate Change*, 8(9), 803–807.
- Harris, D., Foufoula-Georgiou, E., Droegemeier, K. K., & Levit, J. J. (2001). Multiscale statistical properties of a high-resolution precipitation forecast. *Journal of Hydrometeorology*, 2(4), 406–418.
- Harris, L., McRae, A. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). A generative deep learning approach to stochastic downscaling of precipitation forecasts. *arXiv preprint arXiv:2204.02028*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016, Proceedings, Part IV* (pp. 630–645). Springer. doi: 10.1007/978-3-319-46493-0\_38
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049. doi: 10.1002/qj.3803
- Hess, P., Driike, M., Petri, S., Strnad, F., & Boers, N. (2022, 5). *Physically constrained generative adversarial networks for improving precipitation fields from earth system models*. <https://www.codeocean.com/>. doi: 10.24433/CO.2750913.v1
- Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., ... Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning* (pp. 1989–1998).
- IPCC. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (V. Masson-Delmotte et al., Eds.). Cambridge University Press. Retrieved from <https://www.ipcc.ch/report/sixth-assessment-report-working-group-i/> (In Press)
- Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with

- conditional adversarial networks. In *Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* (pp. 5967–5976). doi: 10.1109/CVPR.2017.632
- Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016, Proceedings, Part II* (pp. 694–711). Springer. doi: 10.1007/978-3-319-46475-6\_43
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Krasting, J. P., John, J. G., Blanton, C., McHugh, C., Nikonov, S., Radhakrishnan, A., ... Zhao, M. (2018). *NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 CMIP*. Earth System Grid Federation. doi: 10.22033/ESGF/CMIP6.1407
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.
- Logan, T., Bourgault, P., Smith, T. J., Huard, D., Biner, S., Labonté, M.-P., ... Lavoie, J. (2021, November). *Ouranosinc/xclim: v0.31.0*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.5649661> doi: 10.5281/zenodo.5649661
- Milly, P. C., & Shmakin, A. B. (2002). Global modeling of land water and energy balances. Part I: The land dynamics (LaD) model. *Journal of Hydrometeorology*, *3*(3), 283–299. doi: 10.1175/1525-7541(2002)003<0283:GMOLWA>2.0.CO;2
- Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.
- Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, 193–209.
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*.
- Palmer, T., & Stevens, B. (2019). The scientific challenge of understanding and estimating climate change. *Proceedings of the National Academy of Sciences*, *116*(49), 24390–24395.
- Price, I., & Rasp, S. (2022). Increasing the accuracy and resolution of precipitation forecasts using deep generative models. In *International conference on artificial intelligence and statistics* (pp. 10555–10571).
- Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019, February). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707. doi: 10.1016/j.jcp.2018.10.045
- Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002203.
- Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900.
- Rasp, S., & Thuerey, N. (2021). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, *13*(2), e2020MS002405.
- Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., ... Mohamed, S. (2021). Skillful precipitation nowcasting using deep generative models of radar. *Nature*, *597*, 672–677. doi: 10.1038/s41586-021-03854-z
- Schaphoff, S., Forkel, M., Müller, C., Knauer, J., von Bloh, W., Gerten, D., ... Waha, K. (2018a). LPJmL4 - a dynamic global vegetation model with managed land - Part 1: Model description. *Geoscientific Model Development*, *11*(4), 1343–1375. doi: 10.5194/gmd-11-1343-2018
- Schaphoff, S., Forkel, M., Müller, C., Knauer, J., von Bloh, W., Gerten, D., ... Waha, K. (2018b). LPJmL4 - a dynamic global vegetation model with managed land: Part 2: Model evaluation. *Geoscientific Model Development*, *11*, 1377–1403. doi: 10.5194/gmd-2017-146

- Sinclair, S., & Pegram, G. (2005). Empirical mode decomposition in 2-D space and time: a tool for space-time rainfall analysis and nowcasting. *Hydrology and Earth System Sciences*, 9(3), 127–137.
- Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., . . . Venevsky, S. (2003, 2). Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology*, 9(2), 161–185. doi: 10.1046/j.1365-2486.2003.00569.x
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319–3328).
- Thonicke, K., Spessa, A., Prentice, I. C., Harrison, S. P., Dong, L., & Carmona-Moreno, C. (2010, 6). The influence of vegetation, fire spread and fire behaviour on biomass burning and trace gas emissions: results from a process-based model. *Biogeosciences*, 7(6), 1991–2011. doi: 10.5194/bg-7-1991-2010
- Tian, B., & Dong, X. (2020). The double-itzc bias in cmip3, cmip5, and cmip6 models based on annual mean precipitation. *Geophysical Research Letters*, 47(8), e2020GL087232.
- Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, 12(9), e2019MS002002.
- Tong, Y., Gao, X., Han, Z., Xu, Y., Xu, Y., & Giorgi, F. (2021). Bias correction of temperature and precipitation over China for RCM simulations using the QM and QDM methods. *Climate Dynamics*, 57(5), 1425–1443.
- Traxl, D., Boers, N., Rheinwalt, A., & Bookhagen, B. (2021). The role of cyclonic activity in tropical temperature-rainfall scaling. *Nature communications*, 12(1), 1–9.
- Von Bloh, W., Schaphoff, S., Müller, C., Rolinski, S., Waha, K., & Zaehle, S. (2018). Implementing the nitrogen cycle into the dynamic global vegetation, hydrology, and crop growth model LPJmL (version 5.0). *Geoscientific Model Development*, 11(7), 2789–2812. doi: 10.5194/gmd-11-2789-2018
- Wilcox, E. M., & Donner, L. J. (2007). The frequency of extreme rain events in satellite rain-rate estimates and an atmospheric general circulation model. *Journal of Climate*, 20(1), 53–69.
- Yi, Z., Zhang, H., Tan, P., & Gong, M. (2017, Oct). Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision (iccv)*.
- Zhao, M., Golaz, J.-C., Held, I. M., Guo, H., Balaji, V., Benson, R., . . . Xiang, B. (2018a). The GFDL Global Atmosphere and Land Model AM4.0/LM4.0: 1. Simulation Characteristics With Prescribed SSTs. *Journal of Advances in Modeling Earth Systems*, 10(3), 691–734. doi: 10.1002/2017MS001208
- Zhao, M., Golaz, J.-C., Held, I. M., Guo, H., Balaji, V., Benson, R., . . . Xiang, B. (2018b). The GFDL Global Atmosphere and Land Model AM4.0/LM4.0: 2. Model Description, Sensitivity Studies, and Tuning Strategies. *Journal of Advances in Modeling Earth Systems*, 10(3), 735–769. doi: 10.1002/2017MS001209
- Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).
- Zhuang, J., Dussin, R., Jüling, A., & Rasp, S. (2020, March). *JiaweiZhuang/xESMF: v0.3.0 Adding ESMF.LocStream capabilities*. Zenodo. doi: 10.5281/zenodo.3700105



# Supporting Information for ”Physically Constrained Generative Adversarial Networks for Improving Precipitation Fields from Earth System Models”

Philipp Hess<sup>1,2</sup>, Markus Druke<sup>2</sup>, Stefan Petri<sup>2</sup>, Felix M. Strnad<sup>2,3</sup>, and  
Niklas Boers<sup>1,2,4</sup>

<sup>1</sup>Technical University Munich, Munich, Germany; School of Engineering & Design, Earth System Modelling

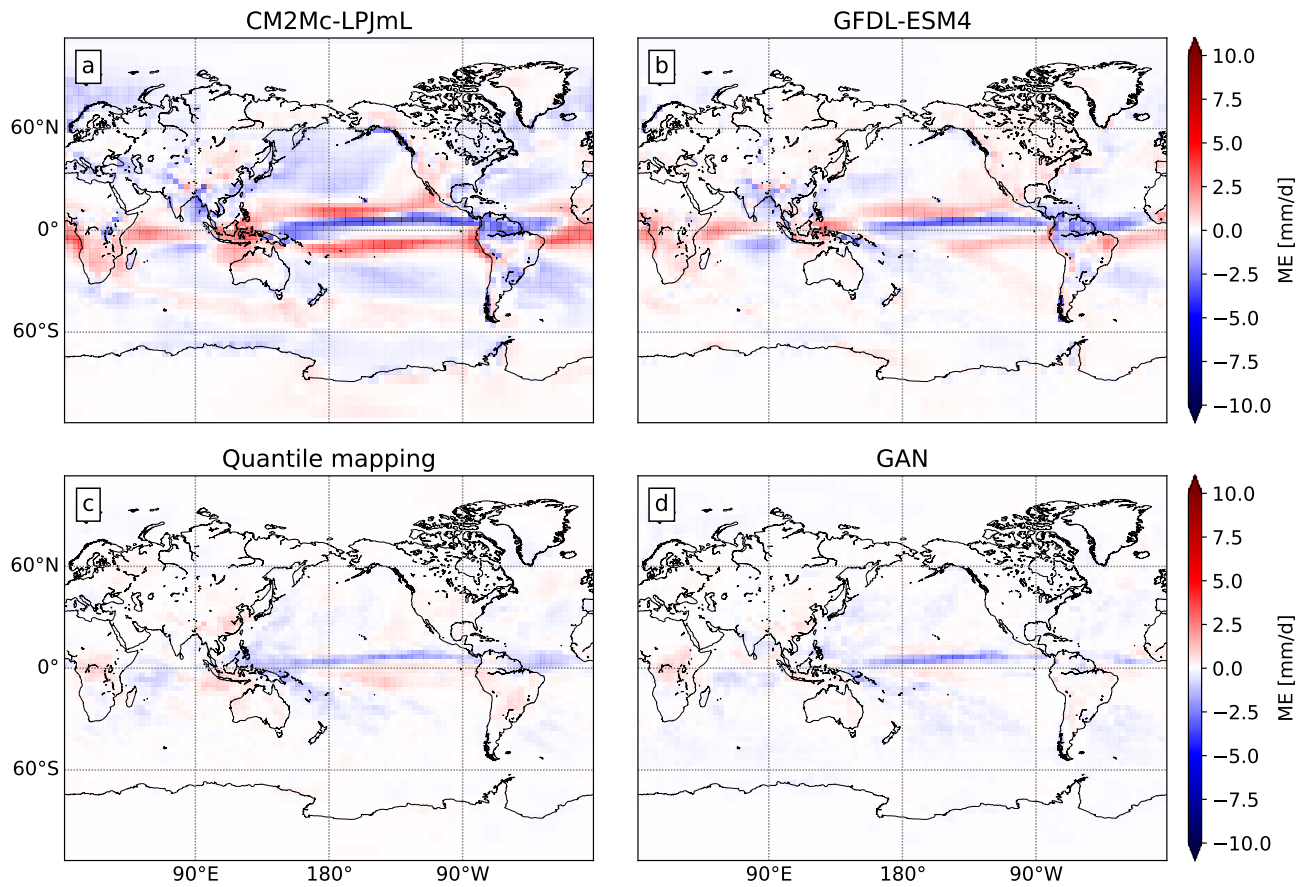
<sup>2</sup>Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany

<sup>3</sup>Cluster of Excellence - Machine Learning for Science, Eberhard Karls Universität Tübingen, Germany

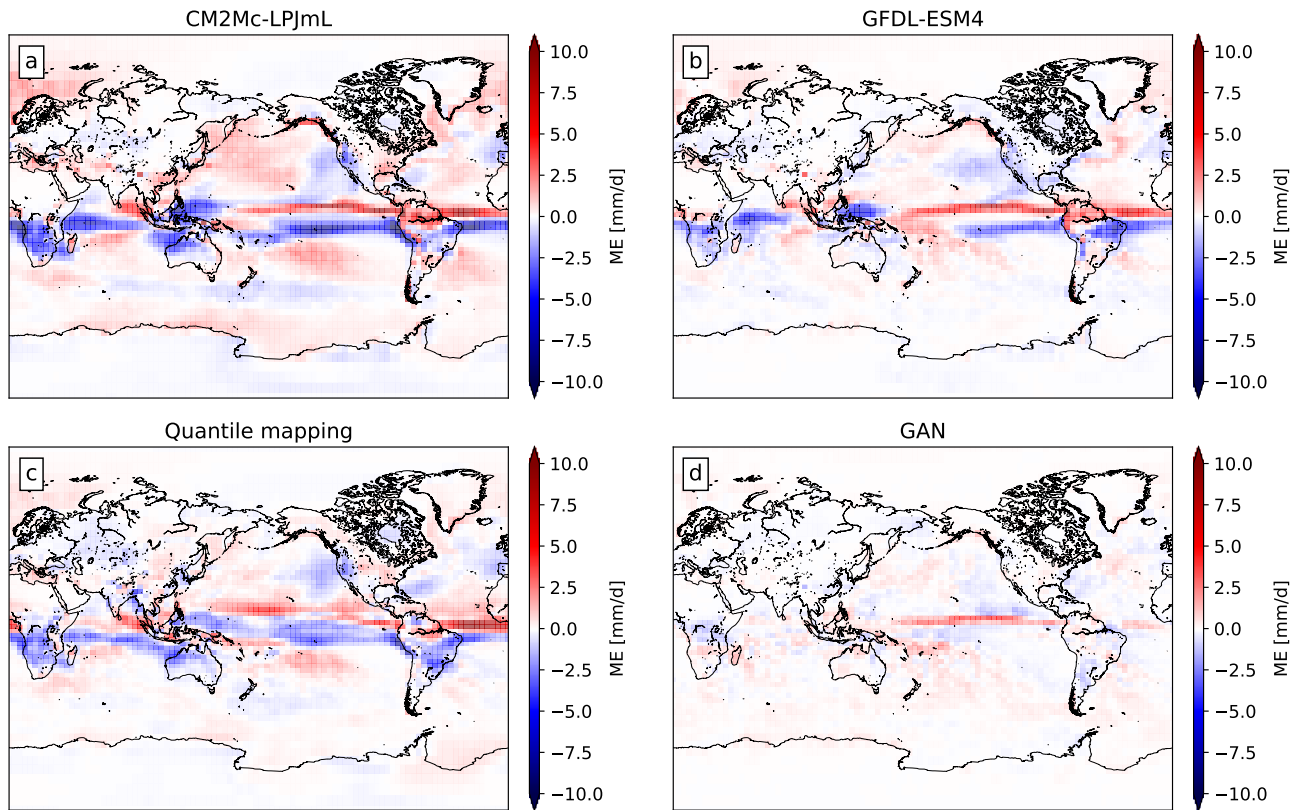
<sup>4</sup>Global Systems Institute and Department of Mathematics, University of Exeter, Exeter, UK

## Contents

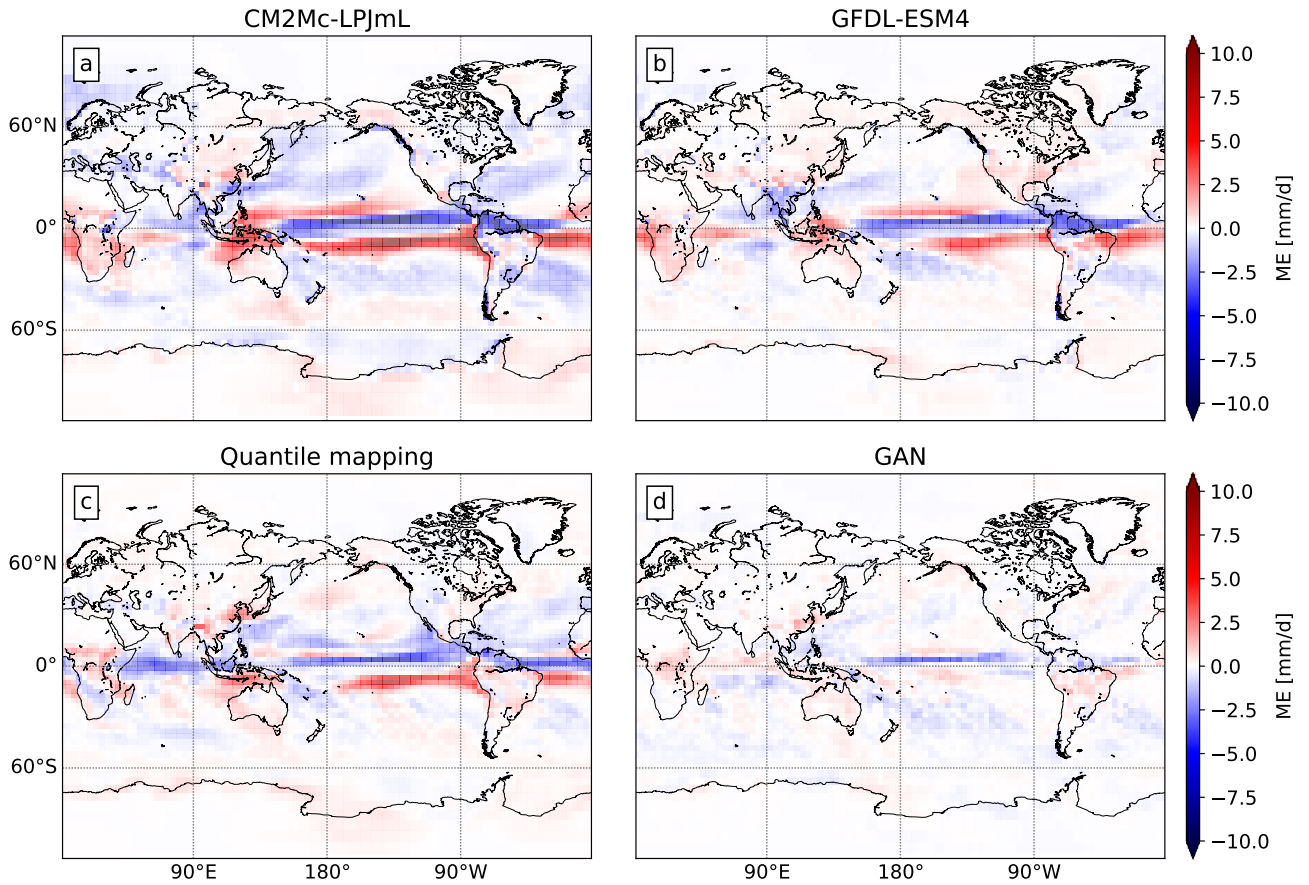
1. Figures S1-S14
2. Table S1-S2



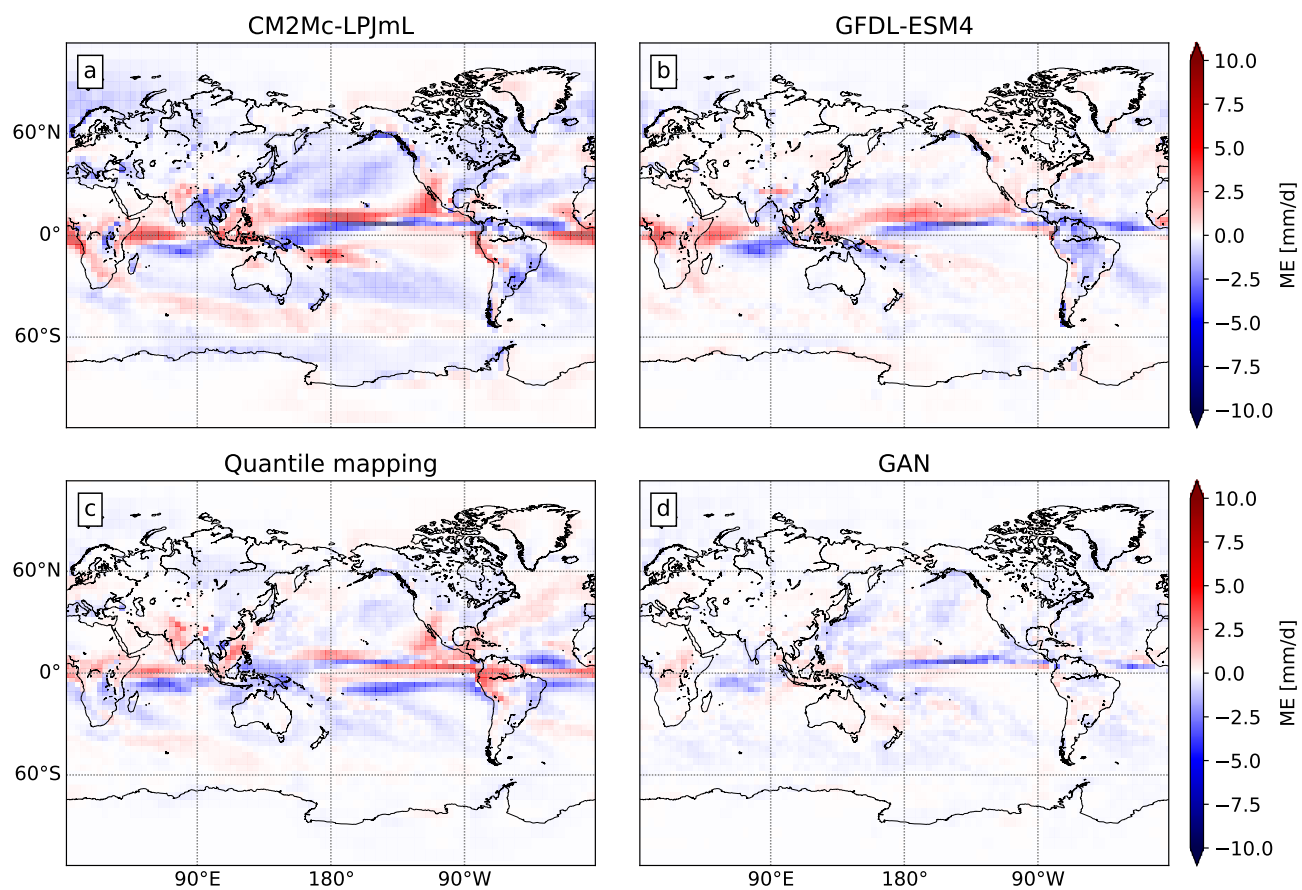
**Figure S1.** Global maps showing the mean error for the entire test set. For (a) CM2Mc-LPJmL, (b) GFDL-ESM4, (c) QM-based and (d) GAN-based post-processing methods applied to the CM2Mc-LPJmL output.



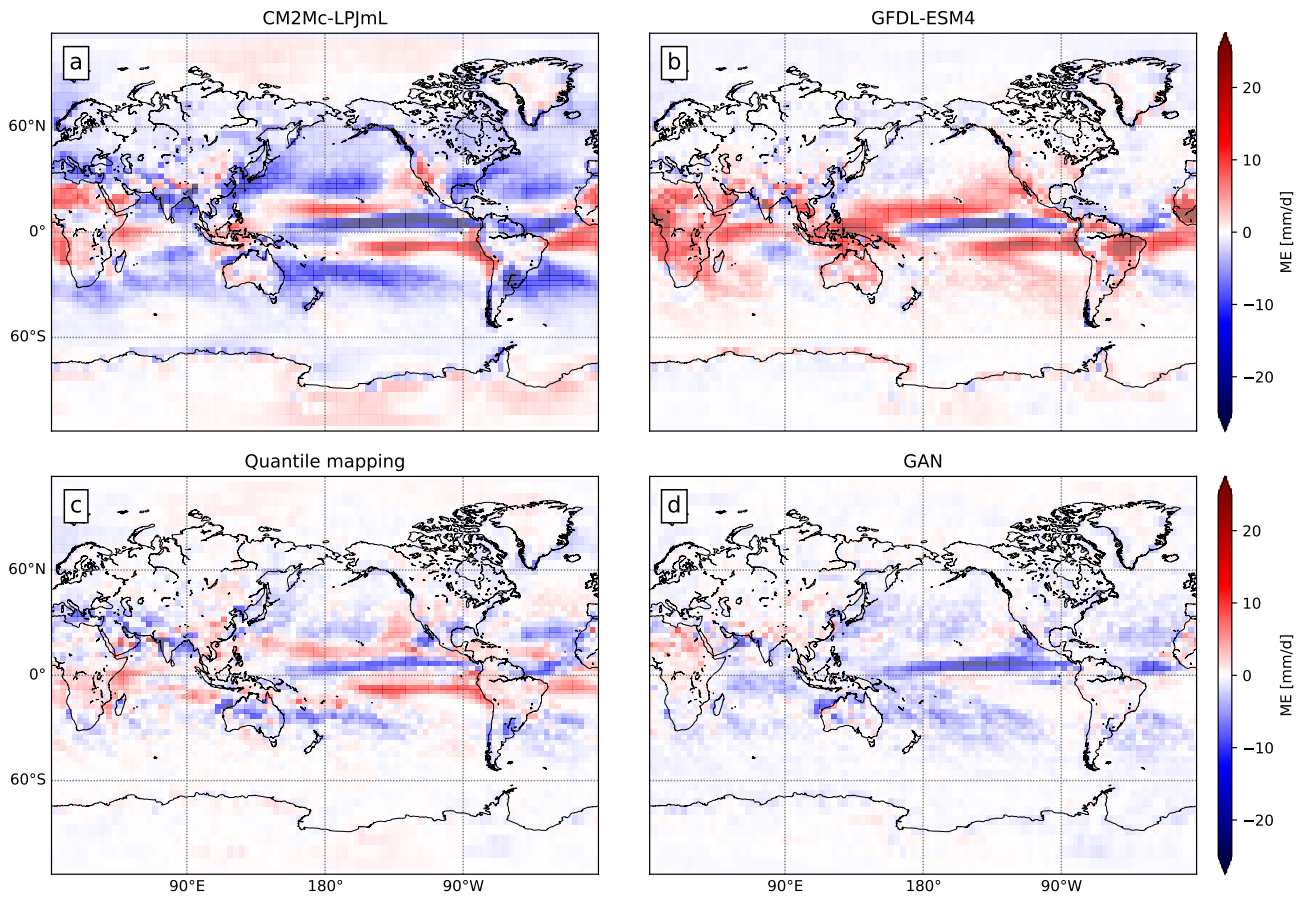
**Figure S2.** Global maps showing the mean error for the DJF season of the test set. For (a) CM2Mc-LPJmL, (b) GFDL-ESM4, (c) QM-based and (d) GAN-based post-processing methods applied to the CM2Mc-LPJmL output.



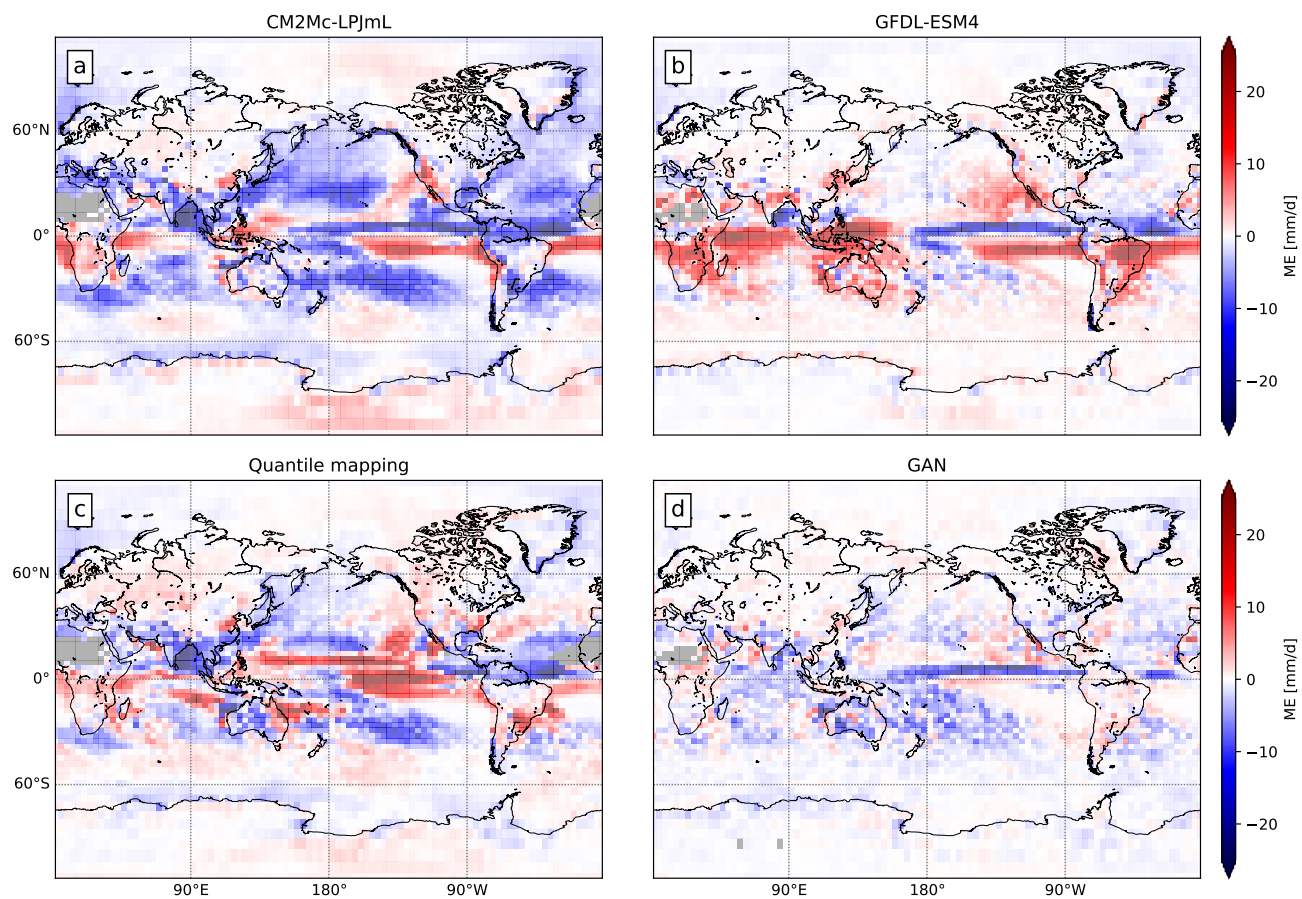
**Figure S3.** Global maps showing the mean error for the MAM season of the test set. For (a) CM2Mc-LPJmL, (b) GFDL-ESM4, (c) QM-based and (d) GAN-based post-processing methods applied to the CM2Mc-LPJmL output.



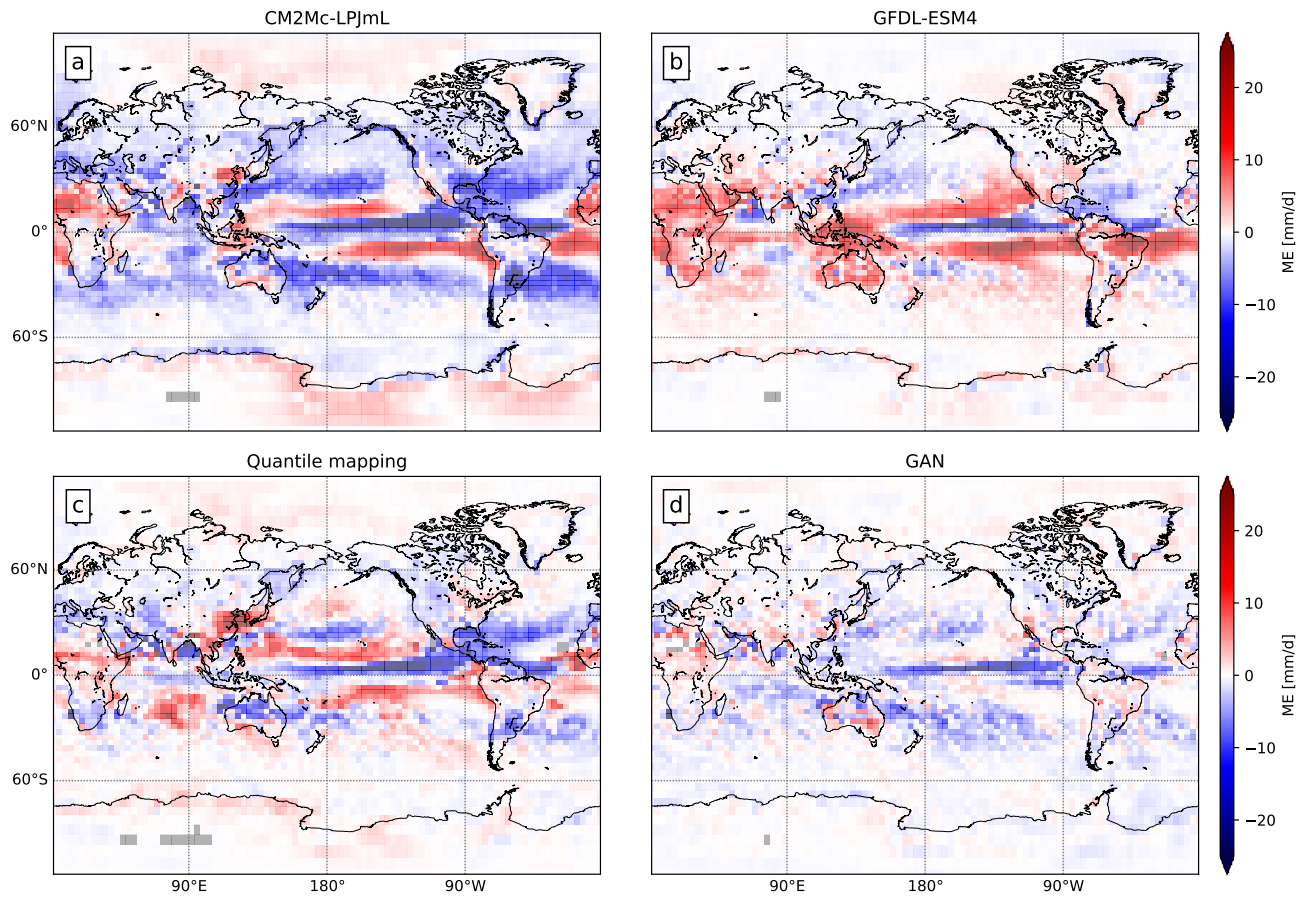
**Figure S4.** Global maps showing the mean error for the SON season of the test set. For (a) CM2Mc-LPJmL, (b) GFDL-ESM4, (c) QM-based and (d) GAN-based post-processing methods applied to the CM2Mc-LPJmL output.



**Figure S5.** Global maps showing the difference in the 95th precipitation percentile for the annual time series of the test set. For (a) CM2Mc-LPJmL, (b) GFDL-ESM4, (c) QM-based and (d) GAN-based post-processing methods applied to the CM2Mc-LPJmL output. Grid cells where the percentiles could not be determined due to insufficient statistics are shown in grey.

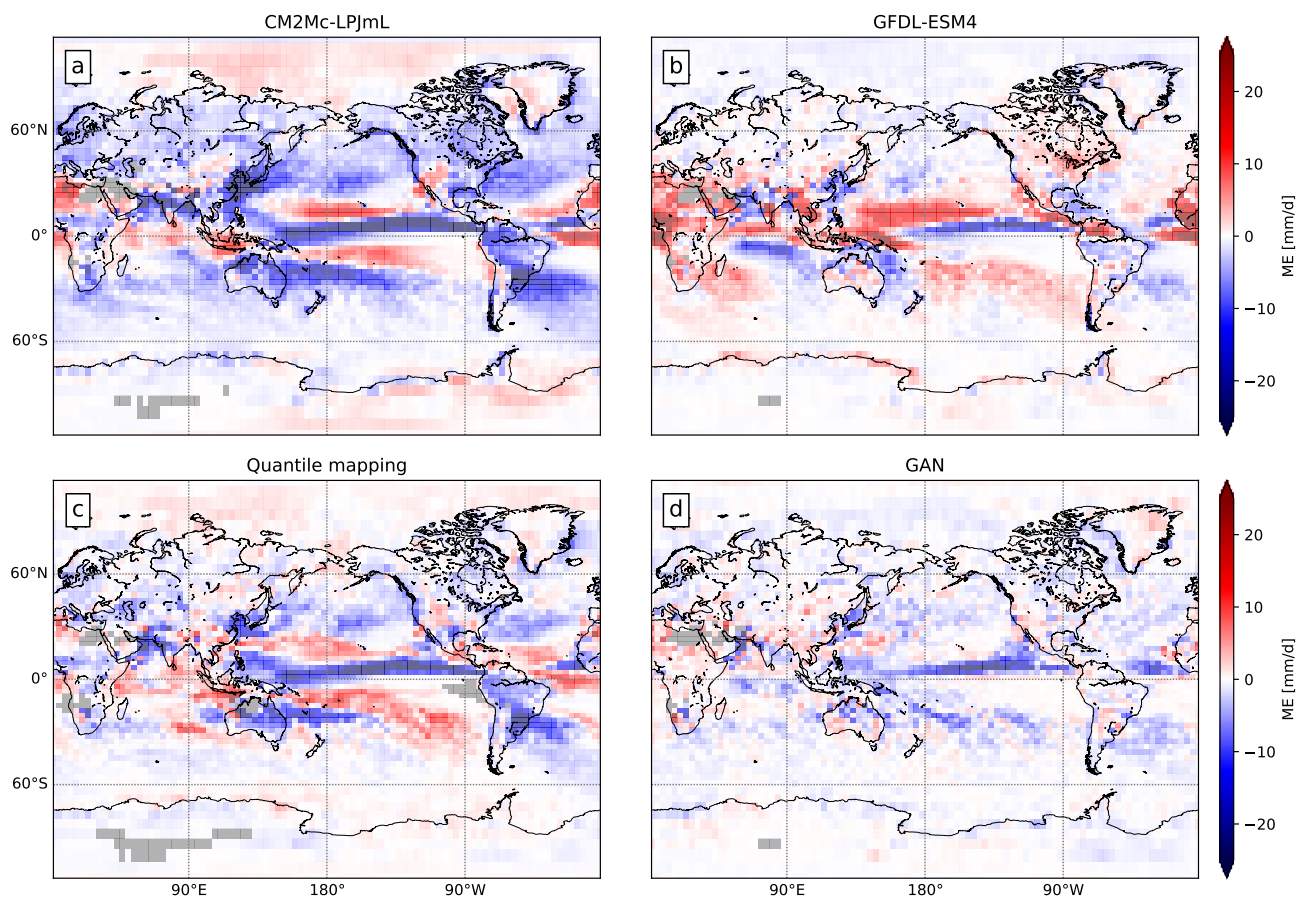


**Figure S6.** Global maps showing the difference in the 95th precipitation percentile for the DJF season of the test set. For (a) CM2Mc-LPJmL, (b) GFDL-ESM4, (c) QM-based and (d) GAN-based post-processing methods applied to the CM2Mc-LPJmL output. Grid cells where the percentiles could not be determined due to insufficient statistics are shown in grey.

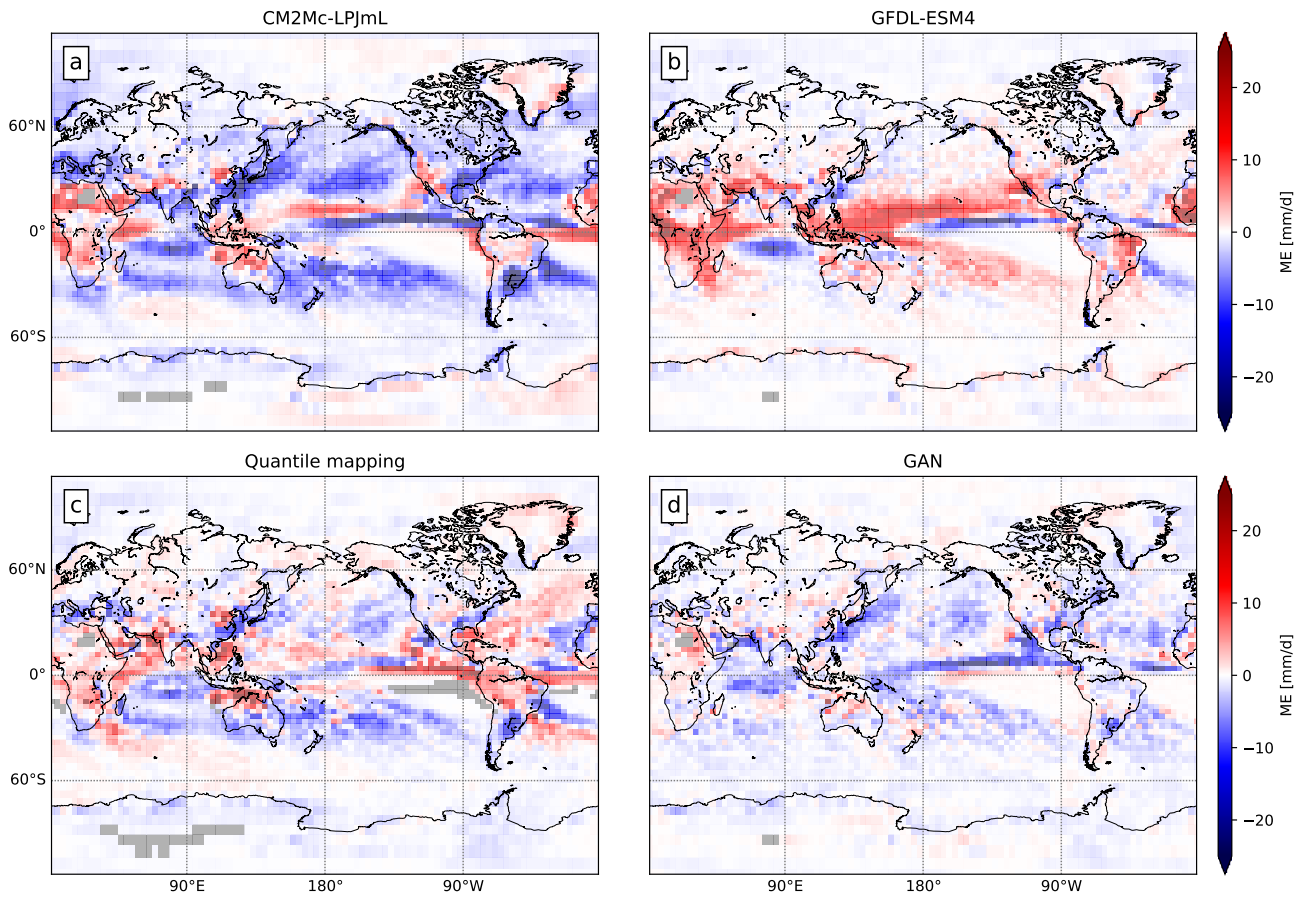


**Figure S7.** Global maps showing the difference in the 95th precipitation percentile for the MAM season of the test set. For (a) CM2Mc-LPJmL, (b) GFDL-ESM4, (c) QM-based and (d) GAN-based post-processing methods applied to the CM2Mc-LPJmL output. Grid cells where the percentiles could not be determined due to insufficient statistics are shown in grey.

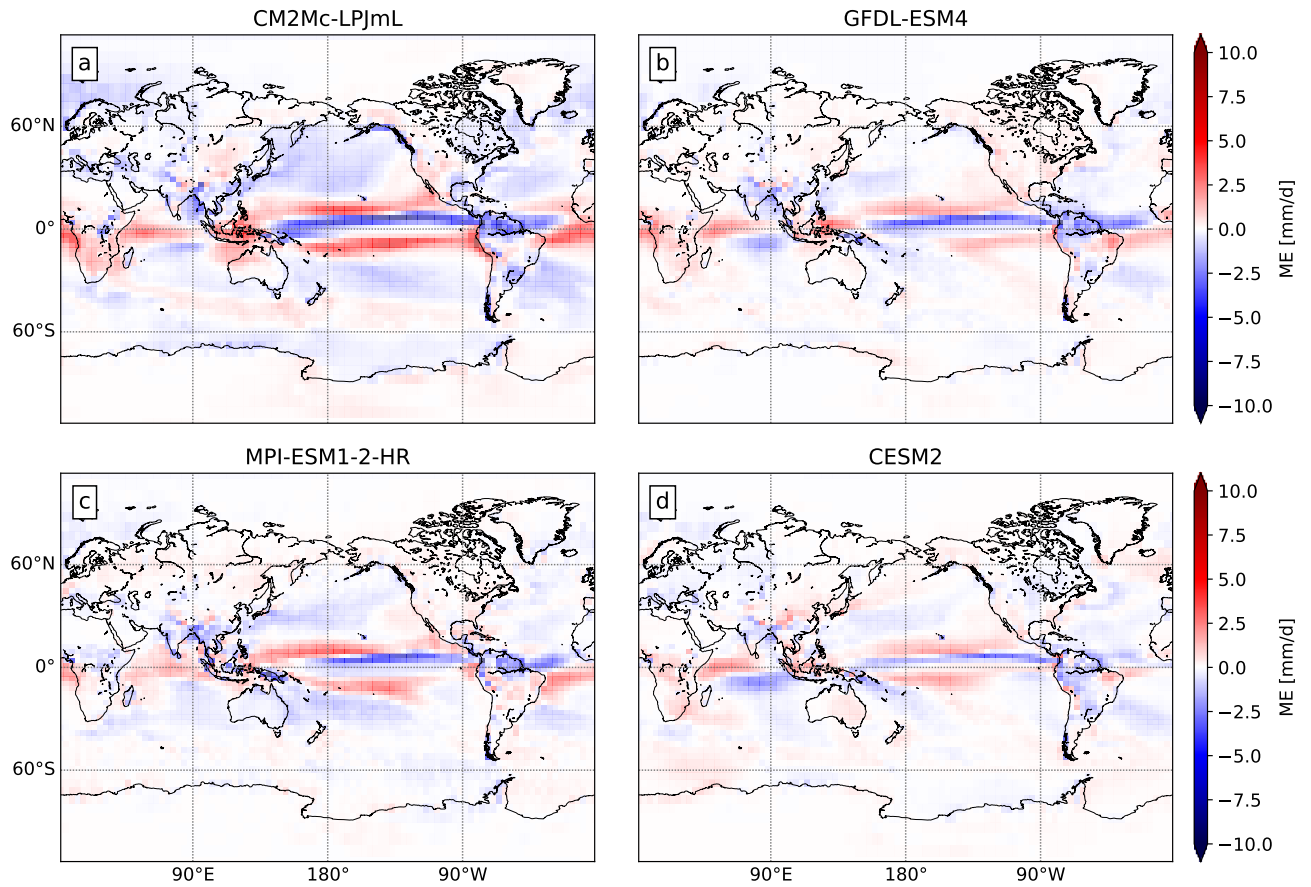




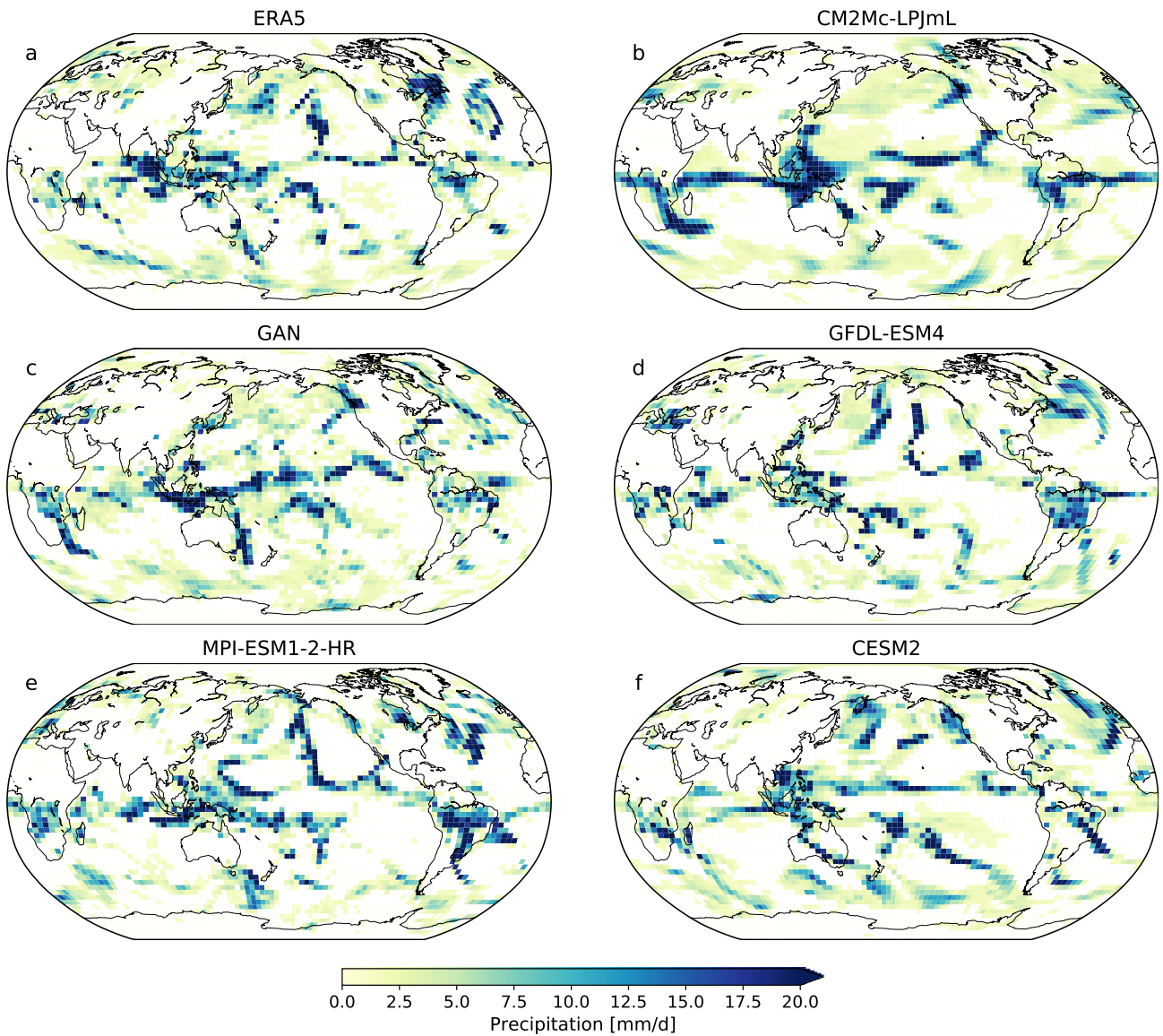
**Figure S8.** Global maps showing the difference in the 95th precipitation percentile for the JJA season of the test set. For (a) CM2Mc-LPJmL, (b) GFDL-ESM4, (c) QM-based and (d) GAN-based post-processing methods applied to the CM2Mc-LPJmL output. Grid cells where the percentiles could not be determined due to insufficient statistics are shown in grey.



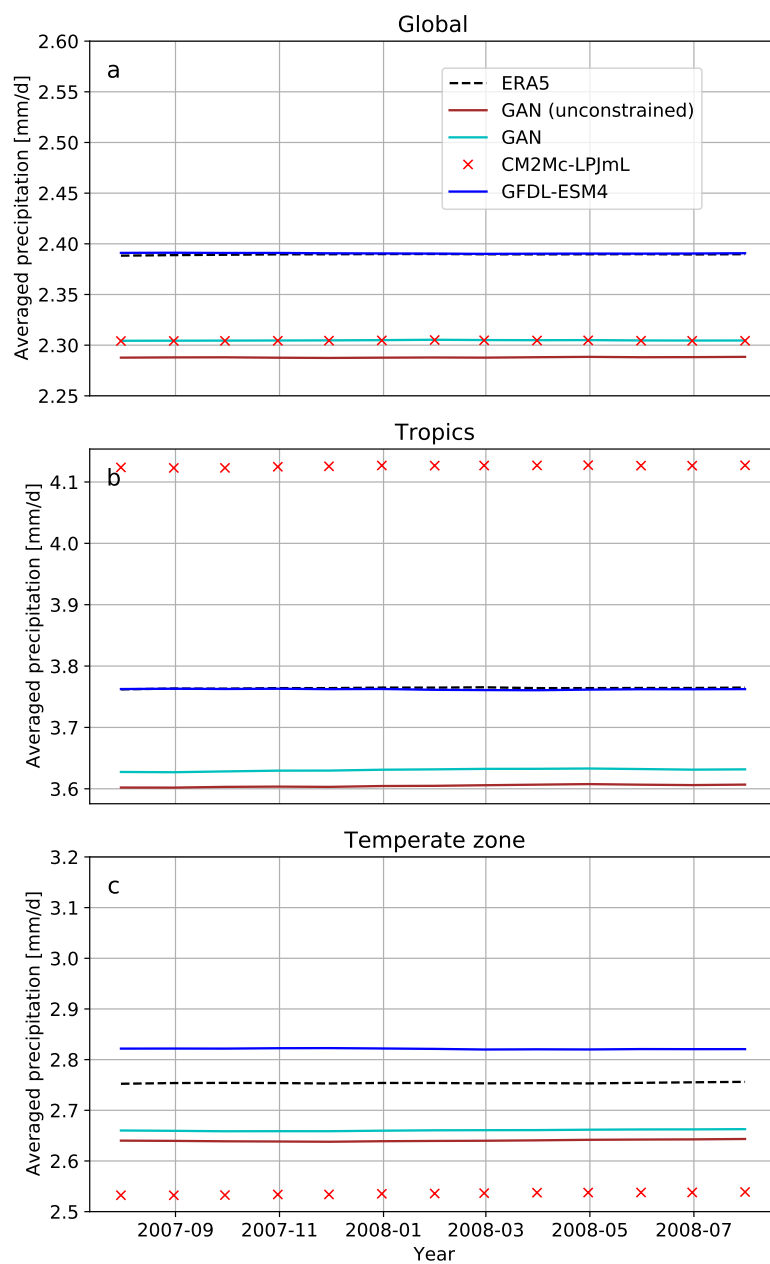
**Figure S9.** Global maps showing the difference in the 95th precipitation percentile for the SON season of the test set. For (a) CM2Mc-LPJmL, (b) GFDL-ESM4, (c) QM-based and (d) GAN-based post-processing methods applied to the CM2Mc-LPJmL output. Grid cells where the percentiles could not be determined due to insufficient statistics are shown in grey.



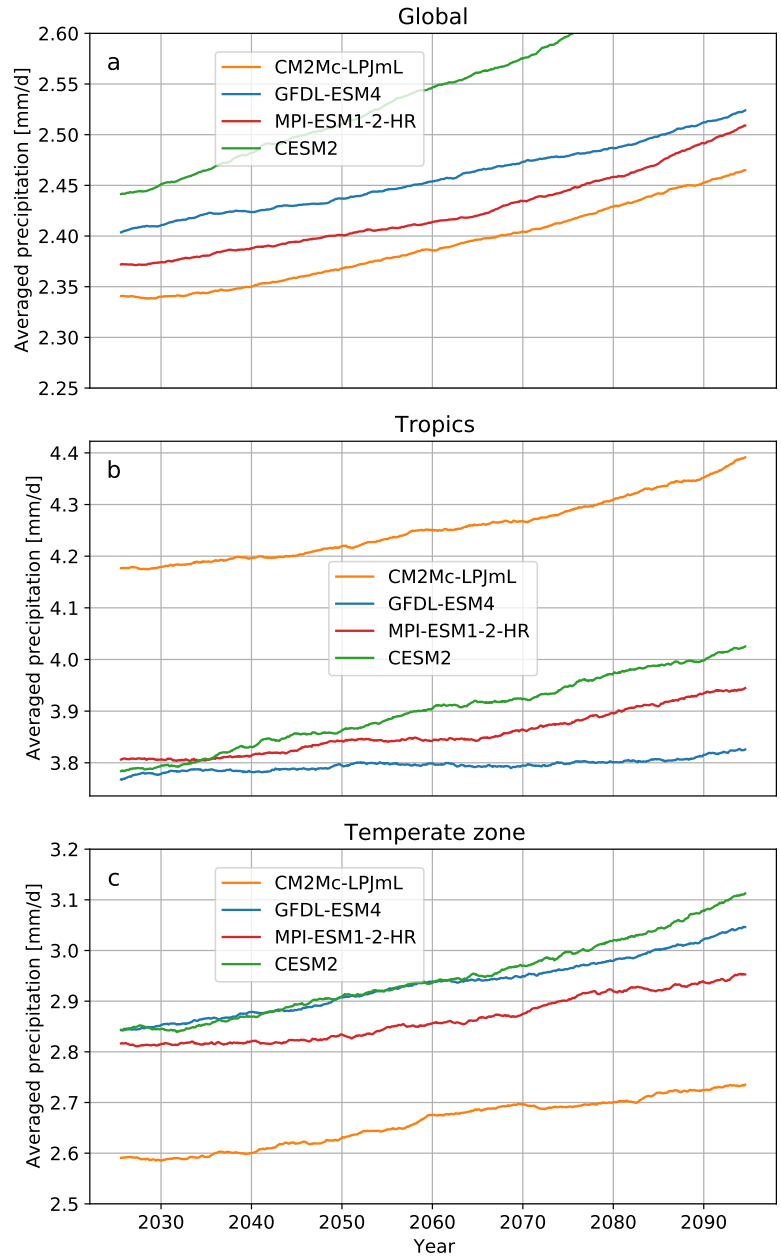
**Figure S10.** Global maps showing the mean error for the annual time series of the test set. For (a) CM2Mc-LPJmL, (b) GFDL-ESM4, (c) MPI-ESM1-2-HR and (d) CESM2.



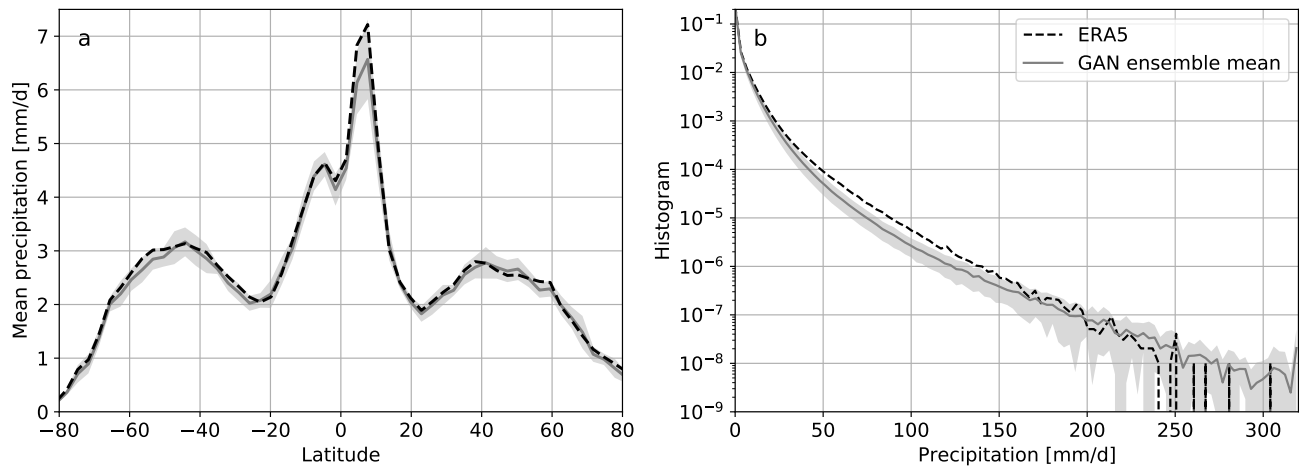
**Figure S11.** Qualitative and quantitative comparison of the intermittency in daily precipitation above 1 mm/day, on the same date (25th December 2014), for the (a) ERA5 reanalysis, (b) CM2Mc-LPJmL model, (c) GAN-based post-processing, (d) GFDL-ESM4, (e) MPI-ESM1-1-HR and (f) CESM2.



**Figure S12.** Large-scale trends as a three year rolling-mean of monthly and spatially average precipitation for the test set period. For (a) global data, (b) the tropics and (c) temperate zone, of the ERA5 reanalysis (black dotted line), CM2Mc-LPJmL (red crosses) and GFDL-ESM4 (blue) models, as well as the constrained (cyan) and unconstrained (brown) GANs.



**Figure S13.** Large-scale precipitation trends are shown for the CMIP6 SSP5-8.5 scenario for the global time series (a), the tropics and temperate zone (c), of the CM2Mc-LPJmL (orange), GFDL-ESM4 (blue), MPI-ESM1-1-HR (red) and CESM2 (green) model.



**Figure S14.** Long-term precipitation statistics based on latitude-profiles and relative frequency histograms for the ERA5 reanalysis (black dotted line) and the ensemble mean of ten GANs (grey, standard deviation as shades) with the same hyperparameters but different checkpoints during the training.

**Table S1.** The averaged absolute value of the grid-cell wise bias is shown for the raw model output of CM2Mc-LPJmL, GFDL-ESM4, MPI-ESM1-1-HR and CESM2.

Season	CM2Mc-LPJmL	GFDL-ESM4	MPI-ESM1-2-HR	CESM2
Annual	0.769	0.448	0.516	0.404
DJF	0.915	0.544	0.677	0.530
MAM	0.886	0.603	0.702	0.549
JJA	0.963	0.589	0.649	0.584
SON	0.823	0.508	0.595	0.513

**Table S2.** The averaged absolute error of the grid-cell-wise 95th precipitation percentiles for the raw CM2Mc-LPJmL and GFDL-ESM4 models, as well as for the QM- and GAN-based post-processing, using the CM2Mc-LPJmL output as input.

Season	CM2Mc-LPJmL	GFDL-ESM4	%	QM	%	GAN	%
Annual	3.715	2.774	25.33	1.868	49.72	<b>1.495</b>	<b>59.76</b>
DJF	4.198	3.071	26.85	3.480	17.10	<b>1.889</b>	<b>55.63</b>
MAM	4.200	3.114	25.86	2.954	29.67	<b>1.876</b>	<b>55.34</b>
JJA	4.324	2.995	30.73	3.077	28.84	<b>1.889</b>	<b>56.31</b>
SON	3.875	2.826	27.07	2.818	27.28	<b>1.972</b>	<b>49.11</b>



## **3.2 P3 | Deep Learning for bias-correcting comprehensive high-resolution Earth system models**

Please turn to the next page.

# Deep Learning for bias-correcting comprehensive high-resolution Earth system models

Philipp Hess<sup>1,2</sup>, Stefan Lange<sup>2</sup>, and Niklas Boers<sup>1,2,3</sup>

<sup>1</sup>Earth System Modelling, School of Engineering & Design, Technical University of Munich, Munich, Germany

<sup>2</sup>Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany

<sup>3</sup>Global Systems Institute and Department of Mathematics, University of Exeter, Exeter, UK

## Key Points:

- A generative adversarial network is shown to improve daily precipitation fields from a state-of-the-art Earth system model.
- Biases in long-term temporal distributions are strongly reduced by the generative adversarial network.
- Our network-based approach can be complemented with quantile mapping to further improve precipitation fields.

## Abstract

The accurate representation of precipitation in Earth system models (ESMs) is crucial for reliable projections of the ecological and socioeconomic impacts in response to anthropogenic global warming. The complex cross-scale interactions of processes that produces precipitation are challenging to model, however, inducing potentially strong biases in ESM fields, especially regarding extremes. State-of-the-art bias correction methods only address errors in the simulated frequency distributions locally, at every individual grid cell. Improving unrealistic spatial patterns of the ESM output, which would require spatial context, has not been possible so far. Here, we show that a post-processing method based on physically constrained generative adversarial networks (GANs) can correct biases of a state-of-the-art, CMIP6-class ESM both in local frequency distributions and in the spatial patterns at once. While our method improves local frequency distributions equally well as gold-standard bias-adjustment frameworks it strongly outperforms any existing methods in the correction of spatial patterns, especially in terms of the characteristic spatial intermittency of precipitation extremes.

## 1 Introduction

Precipitation is a crucial climate variable and changing amounts, frequencies, or spatial distributions have potentially severe ecological and socioeconomic impacts. With global warming projected to continue in the coming decades, assessing the impacts of changes in precipitation characteristics is an urgent challenge (Wilcox & Donner, 2007; Boyle & Klein, 2010; IPCC, 2021). Climate impact models are designed to assess the impacts of global warming on, for example, ecosystems, crop yields, vegetation and other land-surface characteristics, infrastructure, water resources, or the economy in general (Kotz et al., 2022), using the output of climate or Earth system models (ESMs) as input. Especially for reliable assessments of the ecological and socioeconomic impacts, accurate ESM precipitation fields to feed the impact models are therefore crucial.

ESMs are integrated on spatial grids with finite resolution. The resolution is limited by the computational resources that are necessary to perform simulations on decadal to centennial time scales. Current state-of-the-art ESMs have a horizontal resolution on the order of 100km, in exceptional cases going down to 50km. Smaller-scale physical processes that are relevant for the generation of precipitation operate on scales below the size of individual grid cells. These can therefore not be resolved explicitly in ESMs and have to be included as parameterizations of the resolved prognostic variables. These include droplet interactions, turbulence, and phase transitions in clouds that play a central role in the generation of precipitation.

The limited grid resolution hence introduces errors in the simulated precipitation fields, leading to biases in short-term spatial patterns and long-term summary statistics. These biases need to be addressed prior to passing the ESM precipitation fields to impact models. In particular, climate impact models are often developed and calibrated with input data from reanalysis data rather than ESM simulations. These reanalyses are created with data assimilation routines and combine various observations with high-resolution weather models. They hence provide a much more realistic input than the ESM simulations and statistical bias correction methods are necessary to remove biases in the ESM simulations output and to make them more similar to the reanalysis data for which the impact models are calibrated. Quantile mapping (QM) is a standard technique to correct systematic errors in ESM simulations. QM estimates a mapping between distributions from historical simulations and observations that can thereafter be applied to future simulations in order to provide more accurate simulated precipitation fields to impact models (Déqué, 2007; Tong et al., 2021; Gudmundsson et al., 2012; Cannon et al., 2015).

State-of-the-art bias correction methods such as QM are, however, confined to address errors in the simulated frequency distributions locally, i.e., at every grid cell individually.

Unrealistic spatial patterns of the ESM output, which would require spatial context, have therefore so far not been addressed by postprocessing methods. For precipitation this is particularly important because it has characteristic high intermittency not only in time, but also in its spatial patterns. Multivariate bias correction approaches have recently been developed, aiming to improve spatial dependencies (Vrac, 2018; Cannon, 2018). However, these approaches are typically only employed in regional studies, as the dimension of the input becomes too large for global high-resolution ESM simulations. Moreover, such methods have been reported to suffer from instabilities and overfitting, while differences in their applicability and assumptions make them challenging to use (François et al., 2020).

Here, we employ a recently introduced postprocessing method (Hess et al., 2022) based on a cycle-consistent adversarial network (CycleGAN) to consistently improve both local frequency distributions and spatial patterns of state-of-art high-resolution ESM precipitation fields. Artificial neural networks from computer vision and image processing have been successfully applied to various tasks in Earth system science, ranging from weather forecasting (Weyn et al., 2020; Rasp & Thuerey, 2021) to post-processing (Grönquist et al., 2021; Price & Rasp, 2022), by extracting spatial features with convolutional layers (LeCun et al., 2015). Generative adversarial networks (Goodfellow et al., 2014) in particular have emerged as a promising architecture that produces sharp images that are necessary to capture the high-frequency variability of precipitation (Ravuri et al., 2021; Price & Rasp, 2022; Harris et al., 2022). GANs have been specifically developed to be trained on unpaired image datasets (Zhu et al., 2017). This makes them a natural choice for post-processing the output of climate projections, which – unlike weather forecasts – are not nudged to follow the trajectory of observations; due to the chaotic nature of the atmosphere small deviations in the initial conditions or parameters lead to exponentially diverging trajectories (Lorenz, 1996). As a result, numerical weather forecasts lose their deterministic forecast skill after approximately two weeks at most and century-scale climate simulations do not agree with observed daily weather records. Indeed the task of climate models is rather to produce accurate long-term statistics that to agree with observations.

We apply our CycleGAN approach to correct global high-resolution precipitation simulations of the GFDL-ESM4 model (Krasting et al., 2018) as a representative ESM from the Climate Model Intercomparison Project phase 6 (CMIP6). So far, GANs-based approaches have only been applied to postprocess ESM simulations either in a regional context (François et al., 2021), or to a very-low-resolution global ESM (Hess et al., 2022). We show here that a suitably designed CycleGAN is capable of improving even the distributions and spatial patterns of precipitation fields from a state-of-the-art comprehensive ESM, namely GFDL-ESM4. In particular, in contrast to rather specific existing methods for postprocessing ESM output for climate impact modelling, we will show that the CycleGAN is general and can readily be applied to different ESMs and observational datasets used as ground truth.

In order to assure that physical conservation laws are not violated by the GAN-based postprocessing, we include a suitable physical constraint, enforcing that the overall global sum of daily precipitation values is not changed by the GAN-based transformations; essentially, this assures that precipitation is only spatially redistributed (see Methods). By framing bias correction as an image-to-image translation task, our approach corrects both spatial patterns of daily precipitation fields on short time scales and temporal distributions aggregated over decadal time scales. We evaluate the skill to improve spatial patterns and temporal distributions against the gold-standard ISIMIP3BASD framework (Lange, 2019), which relies strongly on QM.

Quantifying the “realisticness” of spatial precipitation patterns is a key problem in current research (Ravuri et al., 2021). We use spatial spectral densities and the fractal dimension of spatial patterns as a measure to quantify the similarity of intermittent and unpaired precipitation fields. We will show that our CycleGAN is indeed spatial context-aware and strongly improves the characteristic intermittency in spatial precipitation patterns. We

will also show that our CycleGAN combined with a subsequent application of ISIMIP3BASD routine leads to the best overall performance.

## 2 Results

We evaluate our CycleGAN method on two different tasks and time scales. First, the correction of daily rainfall frequency distributions at each grid cell locally, aggregated from decade-long time series. Second, we quantify the ability to improve spatial patterns on daily time scales. Our GAN approach is compared to the raw GFDL-ESM4 model output, as well as to the ISIMIP3BASD methodology applied to the GFDL-ESM4 output.

### 2.1 Temporal distributions

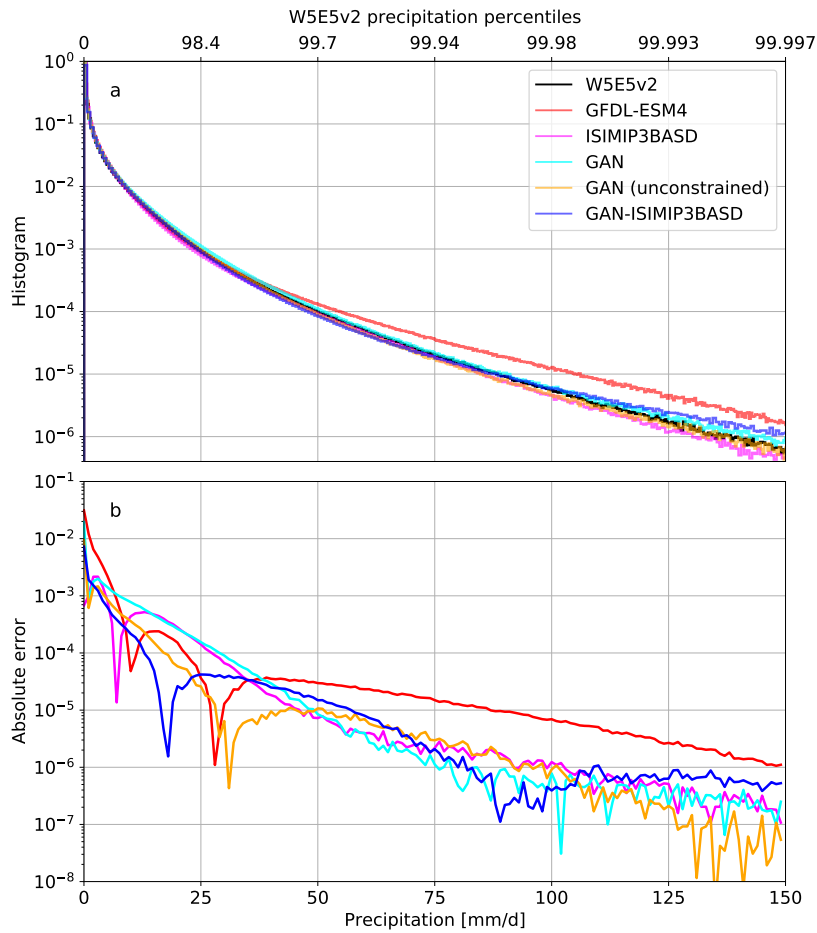


Figure 1: Histograms of relative precipitation frequencies over the entire globe and test period (2004-2014). (a) The histograms are shown for the W5E5v2 ground truth (black), GFDL-ESM4 (red), ISIMIP3BASD (magenta), GAN (cyan), unconstrained GAN (orange), and the constrained-GAN-ISIMIP3BASD combination (blue). (b) Distances of the histograms to the W5E5v2 ground truth are shown for the same models as in (a). Percentiles corresponding to the W5E5v2 precipitation values are given on the second x-axis at the top. Note that GFDL-ESM4 overestimates the frequencies of strong and extreme rainfall events. All compared methods show similar performance in correcting the local frequency distributions.

We compute global histograms of relative precipitation frequencies using daily time series (Fig. 1a). The GFDL-ESM4 model overestimates frequencies in the tail, namely for events above 50 mm/day (i.e., the 99.7th percentile). Our GAN-based method as well as ISIMIP3BASD and the GAN-ISIMIP3BASD combination correct the histogram to match the W5E5v2 ground truth equally well, as can be also seen in the absolute error of the histograms (Fig. 1b).

Comparing the differences in long-term averages of precipitation per grid cell (Fig. 2 and Methods), large biases are apparent in the GFDL-ESM4 model output, especially in the tropics. The double-peaked Intertropical Convergence Zone (ITCZ) bias is visible. The double-ITCZ bias can also be inferred from the latitudinal profile of the precipitation mean in Fig. 3.

Table 1 summarizes the annual biases shown in Fig. 2 as absolute averages, and additionally for the four seasons. The GAN alone reduces the annual bias of the GFDL-ESM4 model by 38.7%. The unconstrained GAN performs better than the physically constrained one, with bias reductions of 50.5%. As expected, the ISIMIP3BASD gives even better results for correcting the local mean, since it is specifically designed to accurately transform the local frequency distributions. It is therefore remarkable that applying the ISIMIP3BASD procedure on the constrained GAN output improves the post-processing further, leading to a local bias reduction of the mean by 63.6%, compared to ISIMIP3BASD with 59.4%. For seasonal time series the order in which the methods perform is the same as for the annual data.

Besides the error in the mean, we also compute differences in the 95th percentile for each grid cell, shown in Fig. S1 and as mean absolute errors in Table 1. Also in this case of heavy precipitation values we find that ISIMIP3BASD outperforms the GAN, but that combining GAN and ISIMIP3BASD leads to best agreement of the locally computed quantiles.

Table 1: The globally averaged absolute value of the grid cell-wise difference in the long-term precipitation average, as well as the 95th percentile, between the W5E5v2 ground truth and GFDL-ESM4, ISIMIP3BASD, GAN, unconstrained GAN, and the GAN-ISIMIP3BASD combination for annual and seasonal time series (in [mm/day]). The relative improvement over the raw GFDL-ESM4 climate model output is shown as percentages for each method.

Season	Percentile	GFDL-ESM4	ISIMIP3-BASD	%	GAN	%	GAN (unconst.)	%	GAN-ISIMIP3-BASD	%
Annual	-	0.535	0.217	59.4	0.328	38.7	0.265	50.5	<b>0.195</b>	<b>63.6</b>
DJF	-	0.634	0.321	49.4	0.395	37.7	0.371	41.5	<b>0.308</b>	<b>51.4</b>
MAM	-	0.722	0.314	56.5	0.419	42.0	0.378	47.6	<b>0.285</b>	<b>60.5</b>
JJA	-	0.743	0.289	61.1	0.451	39.3	0.357	52.0	<b>0.280</b>	<b>62.3</b>
SON	-	0.643	0.327	49.1	0.409	36.4	0.362	43.7	<b>0.306</b>	<b>52.4</b>
Annual	95th	2.264	1.073	52.6	1.415	37.5	1.213	46.4	<b>0.945</b>	<b>58.3</b>
DJF	95th	2.782	1.496	46.2	1.725	38.0	1.655	40.5	<b>1.432</b>	<b>48.5</b>
MAM	95th	2.948	1.482	49.7	1.805	38.8	1.661	43.7	<b>1.337</b>	<b>54.6</b>
JJA	95th	2.944	1.366	53.6	1.852	37.1	1.532	48.0	<b>1.247</b>	<b>57.6</b>
SON	95th	2.689	1.495	44.4	1.741	35.3	1.592	40.8	<b>1.366</b>	<b>49.2</b>

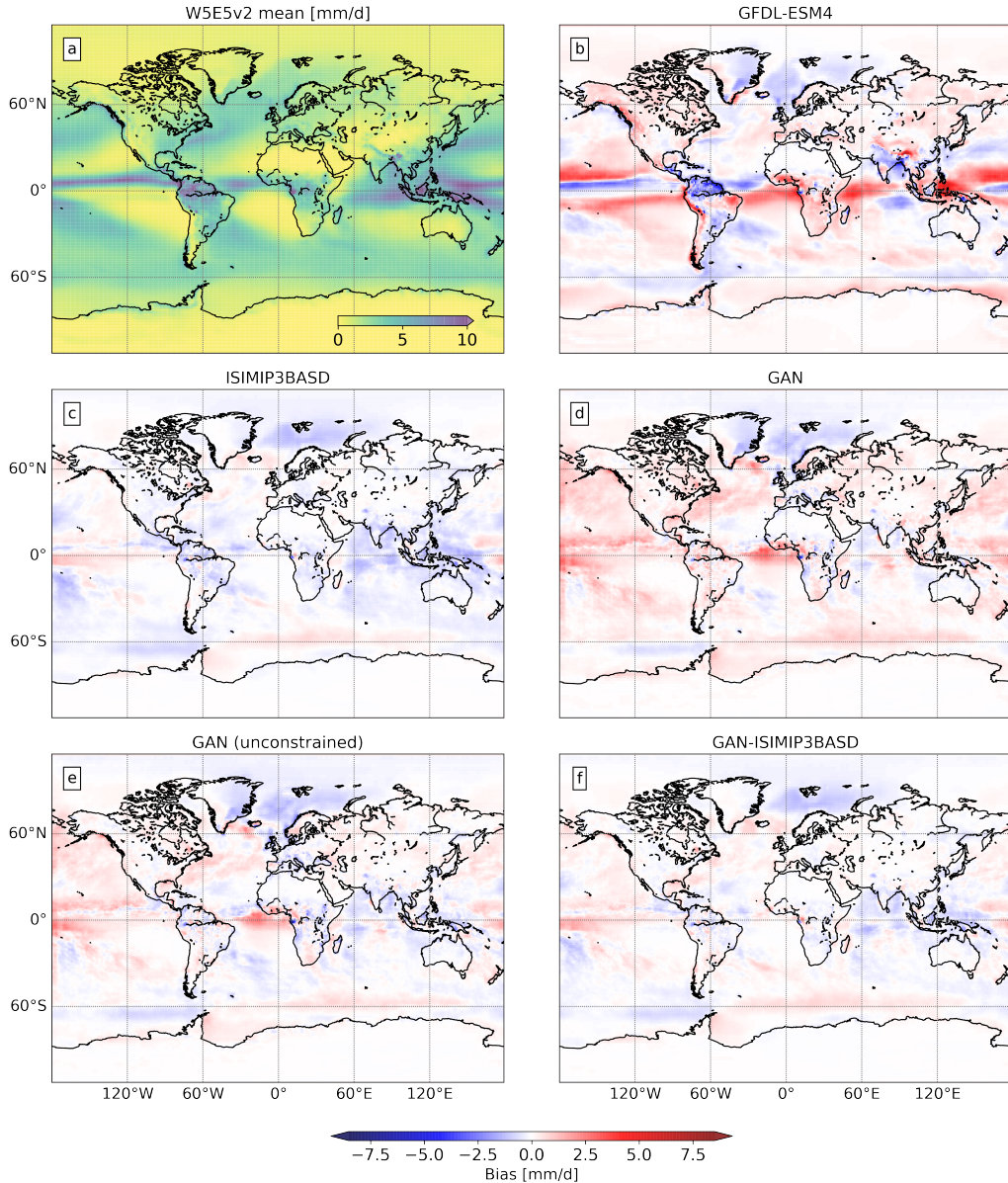


Figure 2: Bias in the long-term average precipitation over the entire test set between the W5E5v2 ground truth (a) and GFDL-ESM4 (b), ISIMIP3BASD (c), GAN (d), unconstrained GAN (e) and the GAN-ISIMIP3BASD combination (f).

## 2.2 Spatial patterns

We compare the ability of the GAN to improve spatial patterns based on the W5E5v2 ground truth, against the GFDL-ESM4 simulations and the ISIMIP3BASD method applied to the GFDL-ESM4 simulations. To model realistic precipitation fields, the characteristic spatial intermittency needs to be captured accurately.

We compute the spatial power spectral density (PSD) of global precipitation fields, averaged over the test set for each method. GFDL-ESM4 shows noticeable deviations from W5E5v2 in the PSD (Fig. 4). Our GAN can correct these over the entire range of wave-

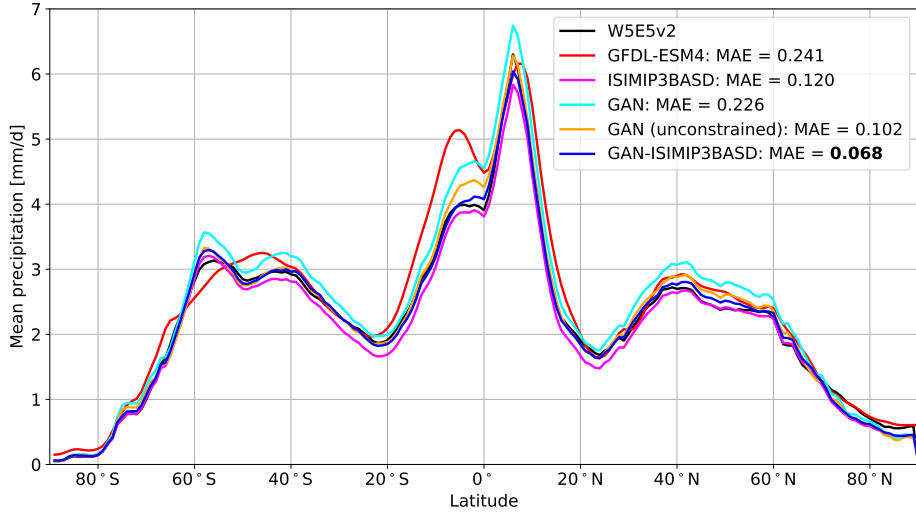


Figure 3: Precipitation averaged over longitudes and the entire test set period from the W5E5v2 ground truth (black) and GFDL-ESM4 (red), ISIMIP3BASD (magenta), GAN (cyan), unconstrained GAN (orange) and the GAN-ISIMIP3BASD combination (blue). To quantify the differences between the shown lines, we show their mean absolute error w.r.t the W5E5v2 ground truth in the legend. These values are different from the ones shown in Table 1 as the average is taken here over the longitudes without their absolute value. The GAN-ISIMIP3BASD approach shows the lowest error.

lengths, closely matching the W5E5v2 ground truth. Improvements over ISIMIP3BASD are especially pronounced in the range of high frequencies (low wavelengths), which are responsible for the intermittent spatial variability of daily precipitation fields. Adding the physical constraint to the GAN does not affect the ability to produce realistic PSD distributions. After applying ISIMIP3BASD to the GAN-processed fields, most of the improvements generated by the GAN are retained, as shown by the GAN-ISIMIP3BASD results.

For a second way to quantifying how realistic the simulated and post-processed precipitation fields are, with a focus on high-frequency spatial intermittency, we investigate the fractal dimension (Edgar & Edgar, 2008) of the lines separating grid cells with daily rainfall sums above and below a given quantile threshold (see Methods). For a sample and qualitative comparison of precipitation fields over the South American continent see Fig. S2. The daily spatial precipitation fields are first converted to binary images using a quantile threshold. The respective quantiles are determined from the precipitation distribution over the entire test set period and globe. The mean of the fractal dimension computed with box-counting (see Methods) (Lovejoy et al., 1987; Meisel et al., 1992; Husain et al., 2021) for each time slice is then investigated (Fig. 5). Both the GFDL-ESM4 simulations themselves and the results of applying the ISIMIP3BASD post-processing to them exhibit spatial patterns with a lower fractal dimension than the W5E5v2 ground truth, implying too low spatial intermittency. In contrast, the GAN translates spatial fields simulated by GFDL-ESM4 in a way that results in closely matching fractal dimensions over the entire range of quantiles.

### 3 Discussion

Postprocessing climate projections is a fundamentally different task from postprocessing weather forecast simulations (Hess et al., 2022). In the latter case, data-driven postprocessing methods, e.g. based on deep learning, to minimize differences between paired samples



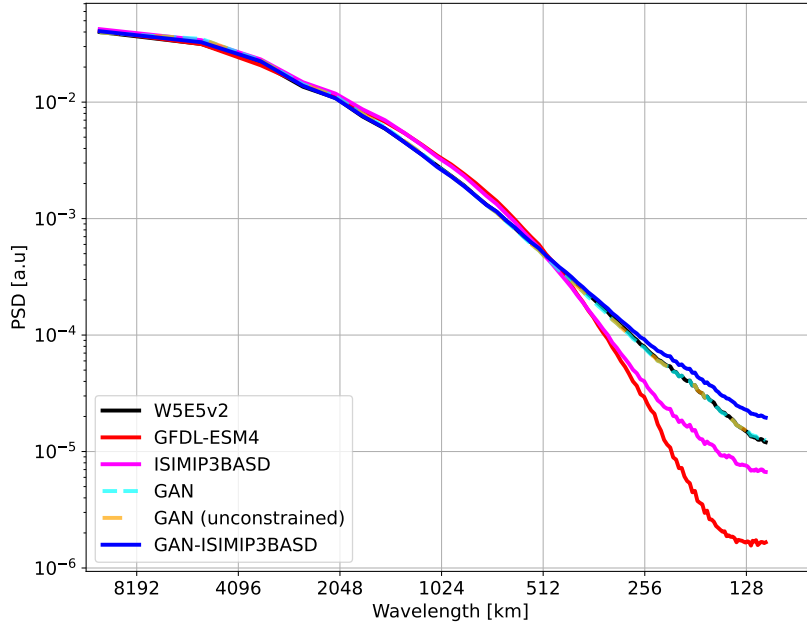


Figure 4: The power spectral density (PSD) of the spatial precipitation fields is shown as an average over all samples in the test set for the W5E5v2 ground truth (black) and GFDL-ESM4 (red), ISIMIP3BASD (magenta), GAN (cyan, dashed), unconstrained GAN (orange, dashed-dotted) and the constrained-GAN-ISIMIP3BASD combination (blue, dotted). The GANs and W5E5v2 ground truth agree so closely that they are indistinguishable. In contrast to ISIMIP3BASD, the GAN can correct the intermittent spectrum accurately over the entire range down to the smallest wavelengths.

of variables such as spatial precipitation fields (Hess & Boers, 2022). Beyond time scales of a few days, however, the chaotic nature of the atmosphere leads to exponentially diverging trajectories, and for climate or Earth system model output there is no observation-based ground truth to directly compare to. We therefore frame the post-processing of ESM projections, with applications for subsequent 195 impact modelling in mind, as an image-to-image translation task with unpaired samples.

To this end we apply a recently developed postprocessing method based on physically constrained CycleGANs to global simulations of a state-of-the-art, high-resolution ESM from the CMIP6 model ensemble, namely the GFDL-ESM4 (Krasting et al., 2018; O' Neill et al., 2016). We evaluate our method against the gold-standard bias correction framework ISIMIP3BASD. Our model can be trained on unpaired samples that are characteristic for climate simulations. It is able to correct the ESM simulations in two regards: temporal distributions over long time scales, including extremes in the distributions' tails, as well as spatial patterns of individual global snap shots of the model output. The latter is not possible with established methods. Our GAN-based approach is designed as a general framework that can be readily applied to different ESMs and observational target datasets. This is in contrast to existing bias-adjustment methods that are often tailored to specific applications.

We chose to correct precipitation because it is arguably one of the hardest variables to represent accurately in ESMs. So far, GANs have only been applied to regional studies or low-resolution global ESMs (François et al., 2021; Hess et al., 2022). The GFDL-ESM4 model simulations are hence chosen in order to test if our CycleGAN approach would lead

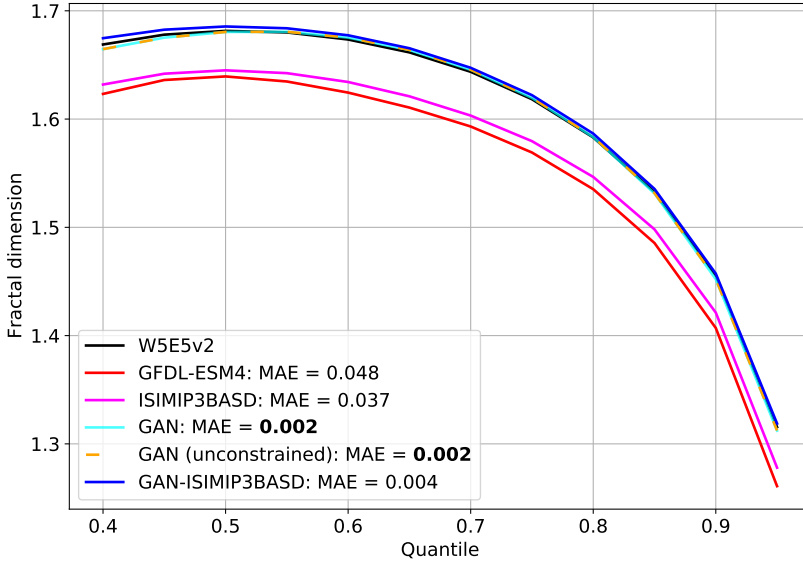


Figure 5: The fractal dimension (see Methods) of binary global precipitation fields is compared as averages for different quantile thresholds. Results are shown for the W5E5v2 ground truth (black) and GFDL-ESM4 (red), ISIMIP3BASD (magenta), GAN (cyan), unconstrained GAN (orange, dashed), and the GAN-ISIMIP3BASD combination (blue). The GAN can accurately reproduce the fractal dimension of the W5E5v2 ground truth spatial precipitation fields over all quantile thresholds, clearly outperforming the ISIMIP3BASD baseline.

to improvements even when postprocessing global high-resolution simulations of one of the most complex and sophisticated ESMS to date. In the same spirit, we evaluate our approach against a very strong baseline given by the state-of-the-art bias correction framework ISIMIP3BASD, which is based on a trend-preserving QM method (Lange, 2019).

Comparing long-term summary statistics, our method yields histograms of relative precipitation frequencies that very closely agree with corresponding histograms from reanalysis data (Fig. 1). The means that the extremes in the far end of the tail are accurately captured, with similar skill to the ISIMIP3BASD baseline that is mainly designed for this task. Differences in the grid cell-wise long-term average show that the GAN skillfully reduces biases (Fig. 2); in particular, the often reported double-peaked ITCZ bias of the GFDL-ESM4 simulations, which is a common feature of most climate models (Tian & Dong, 2020), is strongly reduced (Fig. 3). The ISIMIP3BASD method - being specifically designed for this - produces slightly lower biases for grid-cell-wise averages than the GAN; we show that combining both methods by first applying the GAN and then the ISIMIP3BASD procedure leads to the overall best performance.

Regarding the correction of spatial patterns of the modelled precipitation fields, we compare the spectral density and fractal dimensions of the spatial precipitation fields. Our results show that indeed only the GAN can capture the characteristic spatial intermittency of precipitation closely (Figs. 4 and 5). We believe that the measure of fractal dimension is also relevant for other fields such as nowcasting and medium-range weather forecasting, where blurriness in deep learning-based predictions is often reported (Ravuri et al., 2021) and needs to be further quantified.

Post-processing methods for climate projections have to be able to preserve the trends that result from the non-stationary dynamics of the Earth system on long-time scales. We have therefore introduced the architecture constraint of preserving the global precipitation amount on every day in the climate model output (Hess et al., 2022). We find that this does not affect the quality of the spatial patterns that are produced by our CycleGAN method. However, the skill of correcting mean error biases is slightly reduced by the constraint. This can be expected in part as the constraint is constructed to follow the global mean of the ESM. Hence, biases in the global ESM mean can influence the constrained GAN. This also motivates our choice to demonstrate the combination of the constrained GAN with the QM-based ISIMIP3BASD procedure, since it can be applied to future climate scenarios, making it more suitable for actual applications than the unconstrained architecture.

There are several directions to further develop or approach. The architecture employed here has been built for equally spaced two-dimensional images. Extending the CycleGAN architecture to perform convolutions on the spherical surface, e.g. using graph neural networks, might lead to more efficient and accurate models. Moreover, GANs are comparably difficult to train, which could make it challenging to identify suitable network architectures. Using large ensembles of climate simulations could provide additional training data that could further improve the performance. Another straightforward extension of our method would be the inclusion of further input variables or the prediction additional high-impact physical variables, such as near-surface temperatures that are also important for regional impact models.

## 4 Methods

### 4.1 Training data

We use global fields of daily precipitation with a horizontal resolution of  $1^\circ$  from the GFDL-ESM4 Earth system model (Krasting et al., 2018) and the W5E5v2 reanalysis product (Cucchi et al., 2020; *WFDE5 over land merged with ERA5 over the ocean (W5E5 v2.0)*, 2021) as observation-based ground truth. The W5E5v2 dataset is based on the ERA5 (Hersbach et al., 2020) reanalysis and has been bias-adjusted using the Global Precipitation Climatology Centre (GPCC) full data monthly product v2020 (Schneider et al., 2011) over land and the Global Precipitation Climatology Project (GPCP) v2.3 dataset (Huffman et al., 1997) over the ocean. Both datasets have been regridded to the same  $1^\circ$  horizontal resolution using bilinear interpolation following (Beck et al., 2019). We split the dataset into three periods for training (1950-2000), validation (2001-2003), and testing (2004-2014). This corresponds to 8030 samples for training, 1095 for validation, and 4015 for testing. During pre-processing, the training data is log-transformed with  $\tilde{x} = \log(x + \epsilon) - \log(\epsilon)$  with  $\epsilon = 0.0001$ , following Rasp and Thuerey (2021), to account for zeros in the transform. The data is then normalized to the interval  $[-1, 1]$  following (Zhu et al., 2017).

### 4.2 Cycle-consistent generative adversarial networks

This section gives a brief overview of the CycleGAN used in this study. We refer to (Zhu et al., 2017; Hess et al., 2022) for a more comprehensive description and discussion. Generative adversarial networks learn to generate images that are nearly indistinguishable from real-world examples through a two-player game (Goodfellow et al., 2014). In this set-up, a first network  $G$ , the so-called generator, produces images with the objective to fool a second network  $D$ , the discriminator, which has to classify whether a given sample is generated (“fake”) or drawn from a real-world dataset (“real”). Mathematically this can be formalized as

$$G^* = \min_G \max_D \mathcal{L}_{GAN}(D, G), \quad (1)$$

with  $G^*$  being the optimal generator network. The loss function  $\mathcal{L}_{GAN}(D, G)$  can be defined as

$$\mathcal{L}_{GAN}(D, G) = \mathbb{E}_{y \sim p_y(y)}[\log(D(y))] + \mathbb{E}_{x \sim p_x(x)}[\log(1 - D(G(x)))], \quad (2)$$

where  $p_y(y)$  is the distribution of the real-world target data and samples from  $p_x(x)$  are used as inputs by  $G$  to produce realistic images. The CycleGAN (Zhu et al., 2017) consists of two generator-discriminator pairs, where the generators  $G$  and  $F$  learn inverse mappings between two domains  $X$  and  $Y$ . This allows to define an additional cycle-consistency loss that constraints the training of the networks, i.e.

$$\begin{aligned} \mathcal{L}_{\text{cycle}}(G, F) = & \mathbb{E}_{x \sim p_x(x)}[\|F(G(x)) - x\|_1] \\ & + \mathbb{E}_{y \sim p_y(y)}[\|G(F(y)) - y\|_1]. \end{aligned} \quad (3)$$

It measures the error caused by a translation cycle of an image to the other domain and back. Further, an additional loss term is introduced to regularize the networks to be close to an identity mapping with,

$$\begin{aligned} \mathcal{L}_{\text{ident}}(G, F) = & \mathbb{E}_{x \sim p_x(x)}[\|G(x) - x\|_1] \\ & + \mathbb{E}_{y \sim p_y(y)}[\|F(y) - y\|_1]. \end{aligned} \quad (4)$$

In practice, the log-likelihood loss can be replaced by a mean squared error loss to facilitate a more stable training. Further, the generator loss is reformulated to be minimized by inverting the labels, i.e.

$$\begin{aligned} \mathcal{L}_{\text{Generator}} = & \mathbb{E}_{x \sim p_x(x)}[(D_X(G(x)) - 1)^2] \\ & + \mathbb{E}_{y \sim p_y(y)}[(D_Y(F(y)) - 1)^2] \\ & + \lambda \mathcal{L}_{\text{cycle}}(G, F) + \tilde{\lambda} \mathcal{L}_{\text{ident}}(G, F), \end{aligned} \quad (5)$$

where  $\lambda$  and  $\tilde{\lambda}$  are set to 10 and 5 respectively following (Zhu et al., 2017). The corresponding loss term for the discriminator networks is given by

$$\mathcal{L}_{\text{Discriminator}} = \mathbb{E}_{y \sim p_y(y)}[(D_Y(y) - 1)^2] + \mathbb{E}_{x \sim p_x(x)}[(D_X(G(x)))^2] \quad (6)$$

$$+ \mathbb{E}_{x \sim p_x(x)}[(D_X(x) - 1)^2] + \mathbb{E}_{y \sim p_y(y)}[(D_Y(F(y)))^2]. \quad (7)$$

The weights of the generator and discriminator networks are then optimized with the ADAM (Kingma & Ba, 2014) optimizer using a learning rate of  $2e^{-4}$  and updated in an alternating fashion. We train the network for 350 epochs and a batch size of 1, saving model checkpoints every other epoch. We evaluate the checkpoints on the validation dataset to determine the best model instance.

### 4.3 Network Architectures

Both the generator and discriminator have fully convolutional architectures. The generator uses ReLU activation functions, instance normalization, and reflection padding. The discriminator uses leaky ReLU activations with slope 0.2 instead, together with instance normalization. For a more detailed description, we refer to our previous study (Hess et al., 2022). The network architectures in this study are the same, only with a change in the number of residual layers in the generator network from 6 to 7.

The final layer of the generator can be constrained to preserve the global sum of the input, i.e. by rescaling

$$\tilde{y}_i = y_i \frac{\sum_i^{N_{\text{grid}}} x_i}{\sum_i^{N_{\text{grid}}} y_i}, \quad (8)$$

where  $x_i$  and  $y_i$  are grid cell values of the generator input and output respectively and  $N_{\text{grid}}$  is the number of grid cells. The generator without this constraint will be referred to as unconstrained in this study. The global physical constraint enforces that the global daily precipitation sum is not affected by the CycleGAN postprocessing and hence remains identical to the original value from the GFDL-ESM4 simulations. This is motivated by the observation that large-scale average trends in precipitation follow the Clausius-Clapeyron relation (Traxl et al., 2021), which is based on thermodynamic relations and hence can be expected to be modelled well in GFDL-ESM4.

#### 4.4 Quantile mapping-based bias adjustment

We compare the performance of our GAN-based method to the bias adjustment method ISMIP3BASD v3.0.1 (Lange, 2019, 2022) that has been developed for phase 3 of the Inter-Sectoral Impact Model Intercomparison Project (Warszawski et al., 2014; Frieler et al., 2017). This state-of-the-art bias-adjustment method is based on a trend-preserving quantile mapping (QM) framework. It represents a very strong baseline for comparison as it has been developed prior to this study and used not only in ISMIP3 but also to prepare many of the climate projections that went into the Interactive Atlas produced as part of the 6th assessment report of working group 1 of the Intergovernmental Panel on Climate Change (IPCC, <https://interactive-atlas.ipcc.ch/>). In QM, a transformation between the cumulative distribution functions (CDFs) of the historical simulation and observations is fitted and then applied to future simulations. The CDFs can either be empirical or parametric, the latter being a Bernoulli-gamma distribution for the precipitation in this study. The CDFs are fitted and mapped for each grid cell and day of the year separately. For bias-adjusting the GFDL-ESM4 simulation, parametric QM was found to give the best results, while empirical CDFs are used in combination with the GAN.

To evaluate the methods in this study we define the grid cell-wise bias as the difference in long-term averages as,

$$\text{Bias}(\hat{y}, y) = \frac{1}{T} \sum_{t=1}^T \hat{y}_t - \frac{1}{T} \sum_{t=1}^T y_t, \quad (9)$$

where  $T$  is the number of time steps,  $\hat{y}_t$  and  $y_t$  the modelled and observed precipitation respectively at time step  $t$ .

#### 4.5 Evaluating spatial patterns

Quantifying how realistic spatial precipitation fields are is an ongoing research question in itself, which has become more important with the application of deep learning to weather forecasting and post-processing. In these applications, neural networks often achieve error statistics and skill scores competitive with physical models, while the output fields can at the same time show unphysical characteristics, such as blurring or excessive smoothing. Ravuri et al. (2021) compare the spatial intermittency, which is characteristic of precipitation fields, using the power spectral density (PSD) computed from the spatial fields; in the latter study, the PSD-based quantification was complemented by interviews with a large number of meteorological experts. We propose the fractal dimension of binary precipitation fields as an alternative to quantify how realistic the patterns are.

We compute the fractal dimension via the box-counting algorithm (Lovejoy et al., 1987; Meisel et al., 1992). It quantifies how spatial patterns, for example coastlines (Husain et al., 2021), change with the scale of measurement. The box-counting algorithm divides the image into squares and counts the number of squares that cover the binary pattern of interest,  $N_{\text{squares}}$ . The size of the squares, i.e. the scale of measurement, is then reduced iteratively by a factor  $s$ . The fractal dimension  $D_{\text{fractal}}$  can then be determined from the slope of the resulting log-log scaling, i.e.,

$$D_{\text{fractal}} = \frac{\log(N_{\text{squares}})}{\log(s)}. \quad (10)$$

## Competing interests

The authors declare no competing interests.

## Data availability

The W5E5 data is available for download at <https://doi.org/10.48364/ISIMIP.342217>. The GFDL-ESM4 data can be downloaded at <https://esgf-node.llnl.gov/projects/cmip6/>.

## Code availability

The Python code for processing and analysing the data, together with the PyTorch Lightning (Falcon et al., 2019) code is available at [https://github.com/p-hss/earth\\_system\\_model\\_gan\\_bias\\_correction.git](https://github.com/p-hss/earth_system_model_gan_bias_correction.git). The ISIMIP3BASD code in (Lange, 2022) is used for this study.

## Acknowledgments

NB and PH acknowledge funding by the Volkswagen Foundation, as well as the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research and the Land Brandenburg for supporting this project by providing resources on the high performance computer system at the Potsdam Institute for Climate Impact Research. N.B. acknowledges funding by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 820970 and under the Marie Skłodowska-Curie grant agreement No. 956170, as well as from the Federal Ministry of Education and Research under grant No. 01LS2001A. SL acknowledges funding from the European Union’s Horizon 2022 research and innovation programme under grant agreement no. 101081193 Optimal High Resolution Earth System Models for Exploring Future Climate Changes (OptimESM).

## References

- Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I., ... Adler, R. F. (2019). MSWEP V2 global 3-hourly 0.1° precipitation: methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, 100(3), 473–500. doi: 10.1175/BAMS-D-17-0138.1
- Boyle, J., & Klein, S. A. (2010). Impact of horizontal resolution on climate model forecasts of tropical precipitation and diabatic heating for the TWP-ICE period. *Journal of Geophysical Research: Atmospheres*, 115(D23). doi: 10.1029/2010JD014262
- Cannon, A. J. (2018). Multivariate quantile mapping bias correction: an n-dimensional probability density function transform for climate model simulations of multiple variables. *Climate dynamics*, 50(1), 31–49.
- Cannon, A. J., Sobie, S. R., & Murdock, T. Q. (2015). Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *Journal of Climate*, 28(17), 6938–6959. doi: 10.1175/JCLI-D-14-00754.1
- Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Lange, S., Müller Schmied, H., ... Buontempo, C. (2020). Wfde5: bias-adjusted era5 reanalysis data for impact studies. *Earth System Science Data*, 12(3), 2097–2120. Retrieved from <https://essd.copernicus.org/articles/12/2097/2020/> doi: 10.5194/essd-12-2097-2020

- Déqué, M. (2007). Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global and Planetary Change*, *57*(1-2), 16–26.
- Edgar, G. A., & Edgar, G. A. (2008). *Measure, topology, and fractal geometry* (Vol. 2). Springer.
- Falcon, W., et al. (2019). *PyTorch Lightning*. GitHub repository. Retrieved from <https://github.com/PyTorchLightning/pytorch-lightning>
- François, B., Vrac, M., Cannon, A. J., Robin, Y., & Allard, D. (2020). Multivariate bias corrections of climate simulations: which benefits for which losses? *Earth System Dynamics*, *11*(2), 537–562. Retrieved from <https://esd.copernicus.org/articles/11/537/2020/> doi: 10.5194/esd-11-537-2020
- François, B., Thao, S., & Vrac, M. (2021). Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks. *Climate Dynamics*, *57*(11), 3323–3353.
- Frieler, K., Lange, S., Piontek, F., Reyer, C. P., Schewe, J., Warszawski, L., ... others (2017). Assessing the impacts of 1.5 c global warming–simulation protocol of the inter-sectoral impact model intercomparison project (isimip2b). *Geoscientific Model Development*, *10*(12), 4321–4345.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.
- Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2021). Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200092. doi: 10.1098/rsta.2020.0092
- Gudmundsson, L., Bremnes, J. B., Haugen, J. E., & Engen-Skaugen, T. (2012). Downscaling RCM precipitation to the station scale using statistical transformations—a comparison of methods. *Hydrology and Earth System Sciences*, *16*(9), 3383–3390.
- Harris, L., McRae, A. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). A generative deep learning approach to stochastic downscaling of precipitation forecasts. *arXiv preprint arXiv:2204.02028*.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. doi: 10.1002/qj.3803
- Hess, P., & Boers, N. (2022). Deep learning for improving numerical weather prediction of heavy rainfall. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002765.
- Hess, P., Druke, M., Petri, S., Strnad, F. M., & Boers, N. (2022). Physically constrained generative adversarial networks for improving precipitation fields from earth system models. *Nature Machine Intelligence (accepted)*. Retrieved from <https://arxiv.org/abs/2209.07568> doi: 10.48550/ARXIV.2209.07568
- Huffman, G. J., Adler, R. F., Arkin, P., Chang, A., Ferraro, R., Gruber, A., ... Schneider, U. (1997). The global precipitation climatology project (gpcp) combined precipitation dataset. *Bulletin of the american meteorological society*, *78*(1), 5–20.
- Husain, A., Reddy, J., Bisht, D., & Sajid, M. (2021). Fractal dimension of coastline of australia. *Scientific Reports*, *11*(1), 1–10.
- IPCC. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (V. Masson-Delmotte et al., Eds.). Cambridge University Press. Retrieved from <https://www.ipcc.ch/report/sixth-assessment-report-working-group-i/> (In Press)
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kotz, M., Levermann, A., & Wenz, L. (2022). The effect of rainfall changes on economic production. *Nature*, *601*(7892), 223–227.
- Krasting, J. P., John, J. G., Blanton, C., McHugh, C., Nikonov, S., Radhakrishnan, A.,

- ... Zhao, M. (2018). *NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 CMIP*. Earth System Grid Federation. doi: 10.22033/ESGF/CMIP6.1407
- Lange, S. (2019). Trend-preserving bias adjustment and statistical downscaling with isimip3basd (v1. 0). *Geoscientific Model Development*, 12(7), 3055–3070.
- Lange, S. (2022, June). *Isimip3basd*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.6758997> doi: 10.5281/zenodo.6758997
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lorenz, E. N. (1996). Predictability: A problem partly solved. In *Proc. seminar on predictability* (Vol. 1).
- Lovejoy, S., Schertzer, D., & Tsonis, A. (1987). Functional box-counting and multiple elliptical dimensions in rain. *Science*, 235(4792), 1036–1038.
- Meisel, L., Johnson, M., & Cote, P. (1992). Box-counting multifractal analysis. *Physical Review A*, 45(10), 6989.
- O' Neill, B. C., Tebaldi, C., Van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., ... others (2016). The scenario model intercomparison project (scenariomip) for cmip6. *Geoscientific Model Development*, 9(9), 3461–3482.
- Price, I., & Rasp, S. (2022). Increasing the accuracy and resolution of precipitation forecasts using deep generative models. In *International conference on artificial intelligence and statistics* (pp. 10555–10571).
- Rasp, S., & Thuerey, N. (2021). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, 13(2), e2020MS002405.
- Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., ... Mohamed, S. (2021). Skillful precipitation nowcasting using deep generative models of radar. *Nature*, 597, 672–677. doi: 10.1038/s41586-021-03854-z
- Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., & Ziese, M. (2011). Gpcc full data reanalysis version 6.0 at 0.5: Monthly land-surface precipitation from rain-gauges built on gts-based and historic data. *GPCC Data Rep.*, doi, 10, 585.
- Tian, B., & Dong, X. (2020). The double-itzc bias in cmip3, cmip5, and cmip6 models based on annual mean precipitation. *Geophysical Research Letters*, 47(8), e2020GL087232.
- Tong, Y., Gao, X., Han, Z., Xu, Y., Xu, Y., & Giorgi, F. (2021). Bias correction of temperature and precipitation over China for RCM simulations using the QM and QDM methods. *Climate Dynamics*, 57(5), 1425–1443.
- Traxl, D., Boers, N., Rheinwalt, A., & Bookhagen, B. (2021). The role of cyclonic activity in tropical temperature-rainfall scaling. *Nature communications*, 12(1), 1–9.
- Vrac, M. (2018). Multivariate bias adjustment of high-dimensional climate simulations: the rank resampling for distributions and dependences ( $r^2d^2$ ) bias correction. *Hydrology and Earth System Sciences*, 22(6), 3175–3196. Retrieved from <https://hess.copernicus.org/articles/22/3175/2018/> doi: 10.5194/hess-22-3175-2018
- Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The inter-sectoral impact model intercomparison project (isi-mip): project framework. *Proceedings of the National Academy of Sciences*, 111(9), 3228–3232.
- Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9), e2020MS002109.
- Wfde5 over land merged with era5 over the ocean (w5e5 v2.0)*. (2021). ISIMIP Repository. Retrieved from <https://doi.org/10.48364/ISIMIP.342217> doi: 10.48364/ISIMIP.342217
- Wilcox, E. M., & Donner, L. J. (2007). The frequency of extreme rain events in satellite rain-rate estimates and an atmospheric general circulation model. *Journal of Climate*, 20(1), 53–69.
- Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).



# Supporting Information for ”Deep Learning for bias-correcting comprehensive high-resolution Earth system models”

Philipp Hess<sup>1,2</sup>, Stefan Lange<sup>2</sup>, and Niklas Boers<sup>1,2,3</sup>

<sup>1</sup>Earth System Modelling, School of Engineering & Design, Technical University of Munich, Munich, Germany

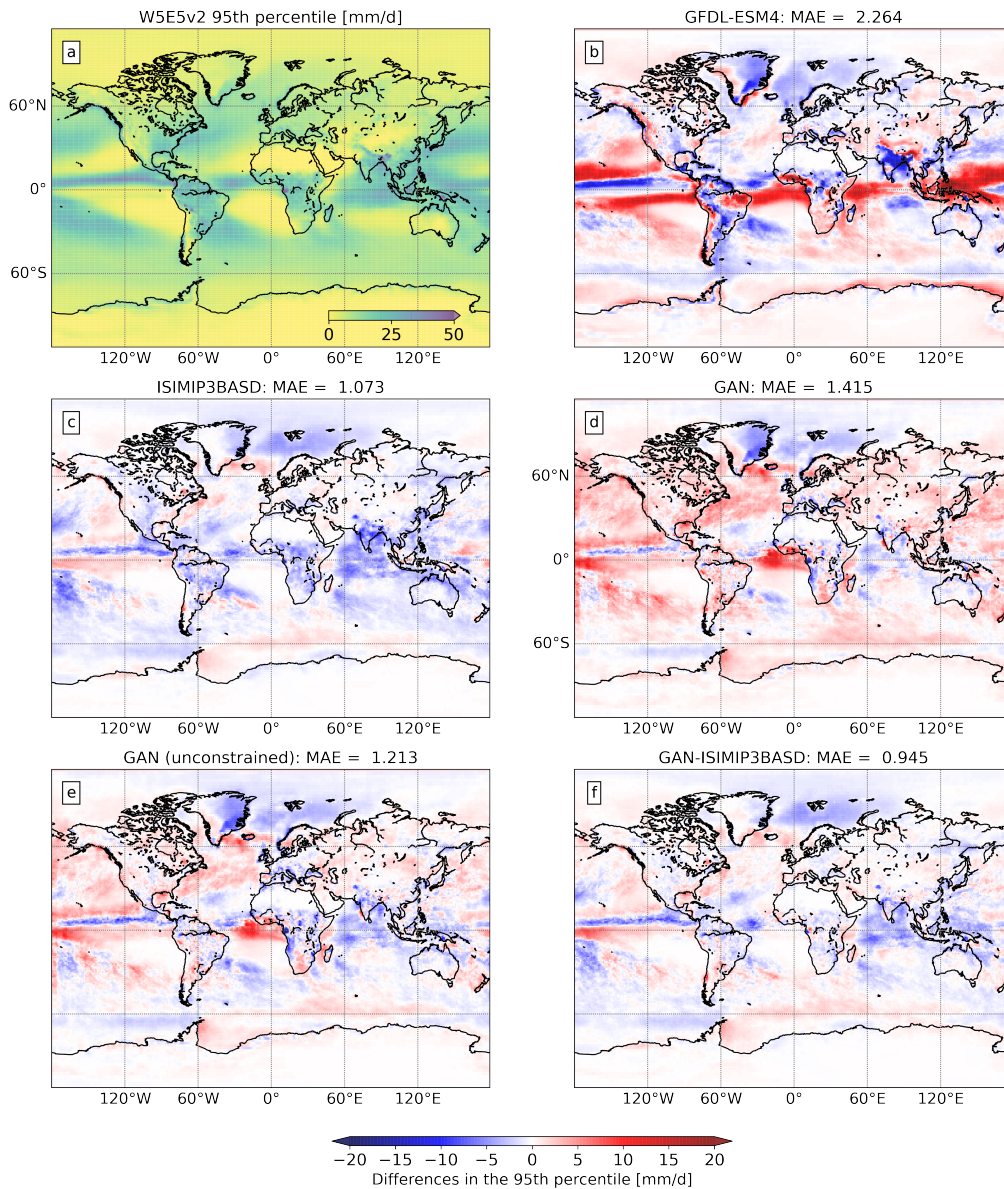
<sup>2</sup>Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany

<sup>3</sup>Global Systems Institute and Department of Mathematics, University of Exeter, Exeter, UK

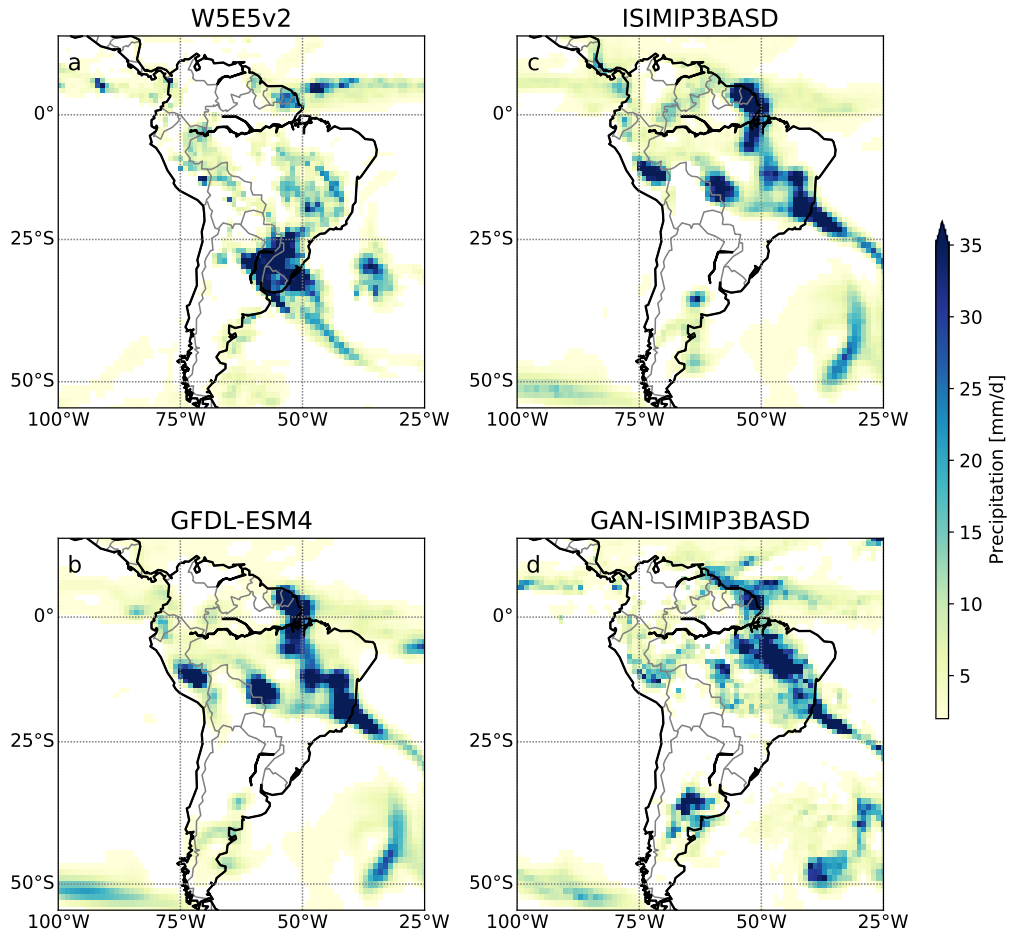
## Contents of this file

1. Figure S1 to S2

---



**Figure S1.** Bias maps as in Fig. 2 but with the 95th percentile instead of the mean. Global mean absolute errors (MAEs) are given in the respective titles. Combining the GAN with ISIMIP3BASD achieves the lowest error compared to the other methods.



**Figure S2.** Qualitative comparison of precipitation fields at the same date (December 21st 2014) over the South American continent. The region is used for a comparison of the fractal dimension in binary precipitation patterns.

### 3. Generative models for improving precipitation fields from climate simulations

## 4 Conclusion

This thesis investigates methods from deep learning in computer vision and image processing to post-processing tasks of precipitation simulations ranging from short-term weather to climate time scales. According to the nature of the training data and post-processing objective, suitable learning strategies are applied.

In weather prediction, the numerical forecast and the observational ground truth follow similar trajectories where the resulting paired training samples allow a supervised learning approach.

For climate projections, the chaotic nature of the atmosphere leads to unpaired samples on short time scales of daily weather. Hence, unsupervised learning approaches for unpaired samples are particularly suitable. Additionally, model interpretability and physical constraints become important for reliable predictions outside the training distribution.

### 4.1 P1 | Post-processing rainfall forecasts with deep neural networks

#### 4.1.1 Main outcomes

Publication P1 developed a deep learning-based post-processing method to improve numerical weather predictions of rainfall, with an emphasis on extreme events. Ensemble mean forecasts with lead times up to 12 hours and three-hourly time steps are taken from the Integrated Forecast System (IFS) ([European Centre for Medium-Range Weather Forecasts, 2012](#)) with near-global coverage and a high spatial resolution of  $0.5^\circ$  horizontally. As ground truth, the satellite-based TRMM 3B42 V7 product ([Huffman et al., 2007](#)) is taken with the same resolution. The relatively high resolution in time and space is particularly important to resolve localized heavy rainfall events.

We apply a convolutional U-Net with a new loss function suitable for the characteristics of spatial rainfall patterns. Regarding the main research questions of this thesis (see section 1.4), the main results in the weather forecasting context are the following.

**Q1 | Improved spatial patterns of precipitation** The loss function for the CNN training includes a multi-scale structural similarity index measure ([Wang et al., 2003](#)), which is sensitive to blurring in images to improve the sharpness of the CNN output. We evaluate the CNN-processed spatial rainfall patterns with a complex wavelet-based extension of the SSIM ([Sampat et al.,](#)

2009) that is invariant to small rotations and translations, i.e., quantifies the similarity of spatial patterns without penalizing small deviations in their position. We find that the CNN-based post-processing leads to improved spatial patterns.

**Q2 | Learning the distribution of precipitation extremes** To learn the strongly skewed distribution of precipitation, a weighted mean-squared error is added to the loss function, where the weights exponentially increase with the rainfall amount. Thus, rare events with large rainfall sums are emphasised during training.

The loss function enables the CNN to correct the frequency distributions over all precipitation values far into the upper distribution tail. A comparable skill to quantile mapping (Déqué, 2007; Cannon et al., 2015), which is specifically designed to correct distributions, is achieved by the CNN.

**Q3 | Increased forecast skill of rare rainfall events** Vertical wind velocities across eleven pressure levels that can be linked to precipitation via updrafts and convection are used as additional input features. The CNN architecture and loss function are suitably chosen for the multi-scale nature of precipitation patterns.

We evaluate the performance with continuous evaluation metrics over the entire range of rainfall values and find that the IFS forecast skill is strongly improved by the CNN-based post-processing. Further, common categorical skill scores are used to evaluate predictions of heavy rainfall events that show an enhancement of the IFS forecast skill by the CNN of factors between 2 to above 6, depending on the percentile threshold. The here-developed method outperforms several baseline ML models and quantile mapping.

### 4.1.2 Context

Post-processing of numerical weather forecasts is very suitable for supervised machine learning tasks, given the large amounts of paired samples available for training. Modern ML methods can efficiently use spatial context, which is not straightforwardly possible with classical statistical methods.

Before emerging ML methods can be integrated into the operational weather forecasting chain, however, accurate and reliable predictions of weather extremes, such as heavy rainfall events, that have a high impact must be demonstrated (Watson, 2022).

To accurately predict extreme and rare rainfall events with neural networks, different approaches have been taken. The authors in Shi et al. (2017); Franch et al. (2020b) use a weighted mean squared error loss function, where the weights are based on selected threshold values. Ebert-Uphoff and Hilburn (2020) use exponential weights similar to our study for estimating radar composite reflectivity from satellite images. Purely neural network-based predictions of temperature extremes have also been found

to benefit from an exponential weighting of the MSE (Lopez-Gomez et al., 2022) following the softmax-based approach of Qi and Majda (2020).

The structural similarity index measure (SSIM) has been used for nowcasting problems (Tran and Song, 2019; Grönquist et al., 2020). Grönquist et al. (2020) apply a supervised U-Net, similar to ours, to ensemble post-processing and find that the network was able to improve predictions of tropical cyclones. Hence, our approach builds on these findings by combining the weighted MSE with multi-scale SSIM in a single loss function.

Recently generative adversarial networks have shown great potential for probabilistic rainfall nowcasting and downscaling (Ravuri et al., 2021; Harris et al., 2022; Price and Rasp, 2022), with the ability to produce sharp forecast fields. In particular, Ravuri et al. (2021) apply a weighted mean error regularization term in the generator loss to ensure that the predictions follow the ground truth closely; however, a degradation of heavy rainfall predictions for longer lead times was found. This highlights the need for further developments in this field with respect to extremes.

### 4.1.3 Outlook

While there have been promising advances in learning extreme weather events with deep neural networks, there remains room for improvements (Watson, 2022).

For nowcasting applications with lead times of minutes to a few hours, purely data-driven methods have shown comparable or better performance with traditional process-based methods (Ravuri et al., 2021). For longer lead times, e.g., medium-range predictions between days to weeks, global numerical models still largely outperform deep learning forecast models (Rasp and Thuerey, 2021; Schultz et al., 2021; Pathak et al., 2022) and some have questioned that the amount of training data will become sufficient for this tasks in the foreseeable future (Rasp and Thuerey, 2021; Palmer, 2022). Therefore, completely replacing the NWP forecast chain with a single neural network “end-to-end” is unlikely to be realized in the near future. A combination of numerical simulation and post-processing will thus likely remain the leading forecast method. Several promising directions exist to build on and extend the work in publication P1.

**Learning extremes** Our research investigates, besides the model architecture and suitable predictors, particularly the design of a loss function to enable the neural network to learn extremes in the target distribution’s tail. Besides using weights in the loss function, different sampling strategies can also be applied to counterbalance the skewed distribution of precipitation, i.e., by giving a higher probability to samples containing extremes (Ravuri et al., 2021). This is particularly suitable for regional forecasts, e.g., over single countries that might not have any or very small rainfall amounts for some of the training samples. In this setting, one could also use training data outside the target region to increase the training sample size. A third

approach to improve the learning of extremes is the development of suitable transformations, e.g., using extreme value theory (Boulaguiem et al., 2022).

**Probabilistic post-processing** Current operational NWP forecast use ensemble runs with perturbed initial conditions to quantify the state-dependent forecast uncertainties (Palmer, 2019). Post-processing can accordingly be extended from deterministic to probabilistic methods to improve the resulting forecast ensemble distribution (Gneiting et al., 2005, 2007). The study in publication P1 so far only corrects the rainfall forecast in a deterministic setting but can, in principle, also be extended to the probabilistic forecasts.

One possibility would be to process each ensemble member forecast individually with the U-Net model and train it with a loss function that evaluates the resulting post-processed ensemble with a scoring rule, such as the Continuous Ranked Probability Score (CRPS), before back-propagating the gradient. Scoring rules such as the CRPS can be extended to include weights emphasising parts, such as tails, of the distribution (Gneiting and Ranjan, 2011; Lerch and Thorarinsdottir, 2013; Lerch et al., 2017). Alternatively, a single network can be used that takes the entire ensemble prediction as input and learns a mapping to an output distribution in the form of a histogram or parametric distribution (Schulz and Lerch, 2022).

Results from Ravuri et al. (2021); Harris et al. (2022) have demonstrated the ability of generative adversarial networks to generate skillful probabilistic forecast distributions. Hence, further research in this direction would also benefit from the findings in publications P2 and P3 that GANs can generate sharp predictions that resolve small-scale precipitation features.

**Evaluation** A thorough evaluation of the trained model is particularly important for extreme events. Watson (2022) provides a summary of evaluation techniques that can be applied in this context, such as quantile-quantile plots, error statistics only for extreme events, and reporting their return periods. An interesting problem in this context is how ML methods respond to extreme events not encountered during the training. Further research is also necessary for quantifying how useful deep learning-based predictions are to end users since the model might achieve state-of-the-art evaluation scores while producing physically unrealistic forecasts at the same time (Ravuri et al., 2021). The question of how to quantify the physical “realisticness” of spatial fields is also relevant in the climate modelling context.

## 4.2 P2 and P3 | Generative models for improving precipitation fields from Earth system simulations



### 4.2.1 Main outcomes

For post-processing precipitation output from Earth system model simulations that run over climate time scales, publications P2 and P3 apply cycle-consistent generative adversarial networks (CycleGANs) (Zhu et al., 2017). Instead of designing a suitable loss function manually, as in P1, the unsupervised CycleGAN learns to improve spatial patterns and distributions in two-player games between generator and discriminator networks. In particular, the ability of CycleGANs to train with unpaired training samples makes them a very suitable choice for processing climate simulations that do not follow the observation-based reanalysis ground truth.

In publication P2, a CycleGAN is trained on daily precipitation fields from the low-resolution CM2Mc-LPJmL ESM (Drüke et al., 2021) and the ERA5 reanalysis (Hersbach et al., 2020) ground truth.

We extend the method developed in P2 with publication P3 in two regards: (i) we apply our method to comprehensive high-resolution GFDL-ESM4 (Krasting et al., 2018) simulations, e.g., that are also used in practice to inform policymakers (IPCC, 2021) and (ii) we compare it to the state-of-the-art post-processing framework ISIMIP3BASD (Lange, 2019) that has been developed outside this study and hence represents a strong baseline for comparisons. We also show in P3 how the strength from both methods - QM and GAN - can be combined to obtain the overall best results.

The main contributions are summarized in the following.

**Q1 | Improved spatial patterns of precipitation** The CycleGAN in P2 and P3 can translate the overly smooth and blurry climate simulations into much sharper fields, s.t. the characteristic intermittency of the spatial patterns is visually indistinguishable from the observation-based reanalysis ground truth.

We evaluate the spatial fields using power spectral densities and, in P3, the fractal dimension that both capture the characteristic small-scale variability and intermittency in spatial precipitation patterns.

The CycleGANs strongly outperform the QM baselines on the correction of spatial patterns.

**Q2 | Learning the distribution of precipitation extremes** The CycleGAN corrects the relative frequency distributions from the ESM s.t. they closely match the ground truth, including extremes in the upper tail.

It achieves comparable or better results than the quantile mapping baseline on low-resolution simulations from the CM2Mc-LPJmL ESM (P2). Using comprehensive high-resolution Earth system simulations (GFDL-ESM4), the same method performs comparably to a state-standard bias correction framework (ISIMIP3BASD) (Lange, 2019). Here, the constrained CycleGAN exhibits larger biases than the unconstrained network while still improving the ESM simulations significantly. Combining the ISIMIP3BASD method with the constrained CycleGAN results in a better performance than either method alone.

**Q4 | Realistic and efficient climate simulations of precipitation** The results from Q1 and Q2 show that CycleGANs are able to make climate simulations much more realistic by correcting spatial patterns on short (daily) time scales and temporal distributions on long (decadal) time scales.

Adding a constraint layer that rescales the output of the CycleGAN to preserve the global ESM precipitation sum per time step is shown in P2 to enable a generalization to out-of-sample predictions in an extreme CMIP6 warming scenario (SSP5-8.5). The constraint ensures consistency with the large-scale hydrological cycle in the simulation and enables the CycleGAN to capture trends in global mean precipitation as expected from thermodynamic considerations (Traxl et al., 2021).

Finally, SmoothGrad (Smilkov et al., 2017), a gradient-based interpretability method, is used in P2 as a sensitivity analysis of the discriminator network. It reveals that a region in the western Pacific Ocean in the input is most important for distinguishing between real and generated precipitation fields. This geographical region also coincides with the largest ESM bias and average precipitation rates, making it a physically reasonable choice (Tian and Dong, 2020).

## 4.2.2 Context

Publications P2 and P3 show the similarity between ESM bias correction and unpaired image-to-image (I2I) translation tasks in deep learning. Concurrently with our study, François et al. (2021) showed that CycleGANs can correct biases in the temperature and precipitation output of a regional climate model over France. The advantage over classical statistical approaches is that spatial patterns can additionally be corrected, besides temporal distributions.

In deep learning, unpaired I2I translation has gained a lot of interest leading to steady improvements over the past years. The CycleGAN architecture (Zhu et al., 2017) was one of the first to demonstrate skillful I2I translation with unpaired samples and established the *cycle-consistency* loss to enforce a bijective mapping between both domains (Yi et al., 2017; Liang et al., 2018; Gokaslan et al., 2018; Tang et al., 2019; Torbunov et al., 2022). A different approach to unpaired I2I translation has been taken with variational autoencoders (VAs) that learn a shared intermediate representation, e.g., with the UNIT model (Liu et al., 2017; Huang et al., 2018). Fulton and Clarke (2021) use the UNIT method to correct biases in an ESM from CMIP6, showing promising results. Diffusion-based extensions have gained in popularity recently and might be a promising alternative to GANs (Sasaki et al., 2021; Zhao et al., 2022).

The interpretability methods for neural networks have also attracted much attention in recent years in the deep learning community (Montavon et al., 2018; Molnar et al., 2020). However, they have so far only been applied to supervised learning tasks in the Earth system science context (to my knowledge) (Ham et al., 2019; McGovern et al., 2019; Toms et al., 2020; Ebert-Uphoff and Hilburn, 2020; Rasp and Thuerey, 2021).

The non-stationarity of the Earth’s climate and the central out-of-sample problem when training machine learning models on historical data are often overlooked. However, the problem has been addressed in sub-grid scale parameterization emulators by enforcing hard constraints on the model architecture (Beucler et al., 2021a) or rescaling target variables to make them invariant to changes in temperature (Beucler et al., 2021b).

### 4.2.3 Outlook

There are several promising directions to extend the CycleGAN-based ESM post-processing in future projects. Extensions to downscaling tasks, different model architectures and ESM integration are briefly outlined in the following.

**Multivariate downscaling** The bias correction is only applied to precipitation in P2 and P3, as it is very challenging to model and because of its large impact. Extending the method to multiple variables should be technically straightforward with convolutional architectures, e.g., by stacking the variables similar to colour channels in RGB images.

Combining the bias correction task with downscaling that increases the ESM resolution is another possible extension, where the unpaired nature of training data has only received limited attention (Saha and Ravela, 2022; Cheng et al., 2022), with the exception of Ballard and Erinjippurath (2022) (to my knowledge). Suppose this approach proves to be successful, and high-resolution fields of multiple variables that are physically consistent can be created with generative models. In that case, data-driven methods might eventually be able to compete with process-based regional climate models on downscaling ESM output. A requirement, however, is the availability of sufficiently high-resolution and comprehensive target training data.

**Model architectures** The CycleGAN has been developed for translating two-dimensional image data, where pixels are equally spaced. The spherical geometry of the Earth, typically represented in terms of latitude-longitude-coordinates, needs to be projected onto the flat image structure. This causes the image pixels to correspond to physical grid cells of changing size depending on the latitude. Thus another direction for future research could be the adoption of convolutional architectures that are more suitable for spherically structured data (Perraudin et al., 2019; Keisler, 2022; Scher and Messori, 2021).

Vision transformer (ViT) networks (Dosovitskiy et al., 2021) have been used to extend the CycleGAN architecture and found to improve the performance for unpaired image-to-image translation (Torbunov et al., 2022). Given the results of the ViT-based FourCastNet (Pathak et al., 2022) for purely data-driven weather predictions, this could be another promising direction to improve the performance of our method.

**ESM integration** Another direction could be integrating the generator network into an ESM as a model component that bias-corrects variables during an ESM run. This could be implemented in two “modes”. The first would serve as an “online” post-processing using the network trained in P2. The second mode would close the feedback loop between GAN and ESM by correcting variables inside the ESM. To increase the stability of the resulting hybrid ESM, the GAN training should be continued during the integration of the hybrid model.

### 4.3 Future developments

To summarize, there are several key challenges in the application of deep neural networks to weather prediction and climate modelling:

1. **Extreme events:** Weather extremes are challenging to model due to the rare occurrence. At the same time, the high impact and potentially devastating consequences of false predictions require reliable forecasts. Hence, deep learning approaches must demonstrate skillful predictions of such events before they can be employed in real-world applications. The work in P1–3 shows how suitable loss functions can enable DNNs to improve the learning of such extremes.
2. **Generalization:** The generalization to out-of-sample predictions is important for weather prediction and especially for climate modelling. In the former, weather events might not be included in the training data due to the limited period where observations are available. In the context of modelling the Earth system under anthropogenic global warming, the non-stationary system leads to out-of-sample predictions when using observational data for training. Since the climate change signal is relatively weak in this period, trends might not be captured accurately. Further, extreme climatic events such as the tipping of entire Earth system components might lead to drastic changes where ESM corrections with respect to historical observations are more difficult to motivate. While this remains an open question, the results of P2 and P3 suggest that suitable constraints show one direction to tackle this out-of-sample problem to capture trends in future climate scenarios.
3. **Physical consistency:** Data-driven models, such as DNNs, should not violate fundamental physical laws, such as conservation of energy and mass, or the positivity of quantities like precipitation. This is particularly important for the generalization to unseen future climates, where a ground truth for evaluation might not be available. Hence, the reliability of the projection is largely based on the physical consistency of the model. Physical constraints, as applied in P2 and P3, can help the neural network to follow conservation laws even in climate regimes unseen during training. Physical consistency is also relevant when

modelling multiple variables. While dynamical process-based models are physically consistent by construction, data-driven methods might fail to learn the inter-variable relation from the data.

4. **Transparency:** To gain insights into whether deep learning models make predictions for the correct physical reasons, interpretability and explainability methods, as applied in P2 can be used. These methods are particularly important for predicting weather extremes and projections to unseen climates, where the reliability of the method is crucial.
5. **Unpaired samples:** Specifically for training DNNs on observational data and long ESM climate runs, unpaired samples that arise from the chaotic nature of the atmosphere pose a challenge for common supervised learning approaches. Since the DNN's prediction and the target ground truth cannot be compared pixel-wise, unsupervised generative models, as applied in P2 and P3, have been found to be a suitable alternative. Besides the DNN training, unpaired model prediction and ground truth data also require new metrics for evaluation, e.g., to assess how realistic spatial patterns are.

The challenges outlined above have so far not been solved entirely. Still, recent work in this field (see previous context sections) and results from this thesis suggest that tools from deep learning are flexible and effective for tackling these. The similarities between domains of deep learning and Earth sciences, such as computer vision and image processing tasks, on the one hand, and post-processing of weather and climate simulations, on the other, offer promising synergies.

Deep learning-based image generation and synthesis have recently gained a lot of attraction, with generative models winning art contests<sup>1</sup>. It thus can be expected that advances in this fast-developing field will help to improve methods tackling similar challenges in ESM post-processing.

By increasing the resolution of ESMs down to kilometer scales where atmospheric convection can be resolved, the need for parameterizing unresolved processes will be reduced (Palmer and Stevens, 2019). However, processes that are crucial for precipitation act on micro-physical scales and hence can not be expected to be explicitly resolved numerically in the near future. Therefore, post-processing will likely be required for future ESM generations, particularly for simulations over climate timescales where the high resolution used in state-of-the-art weather forecasts is computationally prohibitive.

New generations of weather and climate models are being developed in modern programming languages such as Python or Julia (Schneider et al., 2017a; Häfner et al., 2021; Bauer et al., 2021; Ben-Nun et al., 2022), which will improve software interfaces and heterogenous hardware support for combining data-driven methods with numerical process-based models. Therefore,

---

<sup>1</sup><https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html>

the integration of these two approaches can be expected to grow in the future ([Schultz et al., 2021](#); [Irrgang et al., 2021](#)).

# List of figures

1.1	Components of the weather forecasting chain and characteristic time scales of Earth system model components. . . . .	9
1.2	The trade-off between under- and overfitting . . . . .	18





# Bibliography

- Haider Ali, Hayley J. Fowler, and Vimal Mishra. Global Observational Evidence of Strong Linkage Between Dew Point Temperature and Precipitation Extremes. *Geophysical Research Letters*, 45(22):12,320–12,330, 2018. ISSN 1944-8007. doi: 10.1029/2018GL080557. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2018GL080557>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2018GL080557>.
- Troy Arcomano, Istvan Szunyogh, Alexander Wikner, Jaideep Pathak, Brian R. Hunt, and Edward Ott. A Hybrid Approach to Atmospheric Modeling That Combines Machine Learning With a Physics-Based Numerical Model. *Journal of Advances in Modeling Earth Systems*, 14(3):e2021MS002712, 2022. ISSN 1942-2466. doi: 10.1029/2021MS002712. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002712>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021MS002712>.
- Martin Arjovsky and Léon Bottou. Towards Principled Methods for Training Generative Adversarial Networks, January 2017. URL <http://arxiv.org/abs/1701.04862>. arXiv:1701.04862 [cs, stat].
- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223. PMLR, July 2017. URL <https://proceedings.mlr.press/v70/arjovsky17a.html>. ISSN: 2640-3498.
- V. Balaji, Fleur Couvreur, Julie Deshayes, Jacques Gautrais, Frédéric Hourdin, and Catherine Rio. Are general circulation models obsolete? *Proceedings of the National Academy of Sciences*, 119(47):e2202075119, November 2022. doi: 10.1073/pnas.2202075119. URL <https://www.pnas.org/doi/10.1073/pnas.2202075119>. Publisher: Proceedings of the National Academy of Sciences.
- Tristan Ballard and Gopal Erinjippurath. Contrastive Learning for Climate Model Bias Correction and Super-Resolution, November 2022. URL <http://arxiv.org/abs/2211.07555>. arXiv:2211.07555 [physics].
- Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, September 2015. ISSN 14764687. doi: 10.1038/nature14956. Publisher: Nature Publishing Group.
- Peter Bauer, Peter D. Dueben, Torsten Hoefler, Tiago Quintino, Thomas C. Schulthess, and Nils P. Wedi. The digital revolution of Earth-system science. *Nature Computational Science*, 1(2):104–113, 2021. ISSN 2662-8457. doi: 10.1038/s43588-021-00023-0. URL <http://dx.doi.org/10.1038/s43588-021-00023-0>. Publisher: Springer US.
- Hylke E. Beck, Albert I. J. M. van Dijk, Vincenzo Levizzani, Jaap Schellekens, Diego G. Miralles, Brecht Martens, and Ad de Roo. MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, 21(1):589–615, January 2017. ISSN 1027-5606. doi: 10.5194/hess-21-589-2017. URL <https://hess.copernicus.org/articles/21/589/2017/>. Publisher: Copernicus GmbH.
- Hylke E. Beck, Eric F. Wood, Ming Pan, Colby K. Fisher, Diego G. Miralles, Albert I.J.M. Van Dijk, Tim R. McVicar, and Robert F. Adler. MSWep v2 Global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, 100(3): 473–500, 2019. ISSN 00030007. doi: 10.1175/BAMS-D-17-0138.1.

- Tal Ben-Nun, Linus Groner, Florian Deconinck, Tobias Wicky, Eddie Davis, Johann Dahm, Oliver D. Elbert, Rhea George, Jeremy McGibbon, Lukas Trümper, Elynn Wu, Oliver Fuhrer, Thomas Schulthess, and Torsten Hoefler. Productive Performance Engineering for Weather and Climate Modeling with Python, August 2022. URL <http://arxiv.org/abs/2205.04148>. arXiv:2205.04148 [cs].
- Stanley G. Benjamin, John M. Brown, Gilbert Brunet, Peter Lynch, Kazuo Saito, and Thomas W. Schlatter. 100 Years of Progress in Forecasting and NWP Applications. *Meteorological Monographs*, 59(1):13.1–13.67, December 2018. doi: 10.1175/AMSMONOGRAPH-D-18-0020.1. URL <https://journals.ametsoc.org/view/journals/amsm/59/1/amsmonographs-d-18-0020.1.xml>. Publisher: American Meteorological Society Section: Meteorological Monographs.
- Peter Berg, Christopher Moseley, and Jan O. Haerter. Strong increase in convective precipitation in response to higher temperatures. *Nature Geoscience*, 6(3):181–185, March 2013. ISSN 17520894. doi: 10.1038/ngeo1731. URL [www.nature.com/naturegeoscience](http://www.nature.com/naturegeoscience). Publisher: Nature Publishing Group.
- Tom Beucler, Michael Pritchard, Stephan Rasp, Jordan Ott, Pierre Baldi, and Pierre Gentine. Enforcing Analytic Constraints in Neural Networks Emulating Physical Systems. *Physical Review Letters*, 126(9):98302, 2021a. ISSN 10797114. doi: 10.1103/PhysRevLett.126.098302. URL <https://doi.org/10.1103/PhysRevLett.126.098302>. arXiv: 1909.00912 Publisher: American Physical Society.
- Tom Beucler, Michael Pritchard, Janni Yuval, Ankitesh Gupta, Liran Peng, Stephan Rasp, Fiaz Ahmed, Paul A. O’Gorman, J. David Neelin, Nicholas J. Lutsko, and Pierre Gentine. Climate-Invariant Machine Learning, December 2021b. URL <http://arxiv.org/abs/2112.08440>. arXiv:2112.08440 [physics].
- Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Gutttag. What is the State of Neural Network Pruning? *Proceedings of Machine Learning and Systems*, 2:129–146, March 2020. URL <https://proceedings.mlsys.org/paper/2020/hash/d2ddea18f00665ce8623e36bd4e3c7c5-Abstract.html>.
- Niklas Boers, Bedartha Goswami, Aljoscha Rheinwalt, Bodo Bookhagen, Brian Hoskins, and Jürgen Kurths. Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*, 566(7744):373–377, February 2019. ISSN 14764687. doi: 10.1038/s41586-018-0872-x. URL <http://www.nature.com/articles/s41586-018-0872-x>. Publisher: Nature Publishing Group.
- Younes Boulaguiem, Jakob Zscheischler, Edoardo Vignotto, Karin van der Wiel, and Sebastian Engelke. Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks. *Environmental Data Science*, 1:e5, 2022. ISSN 2634-4602. doi: 10.1017/eds.2022.4. URL <https://doi.org/10.1017/eds.2022.4>. Publisher: Cambridge University Press.
- Alex J. Cannon, Stephen R. Sobie, and Trevor Q. Murdock. Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes? *Journal of Climate*, 28(17):6938–6959, September 2015. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-14-00754.1. URL <https://journals.ametsoc.org/view/journals/clim/28/17/jcli-d-14-00754.1.xml>. Publisher: American Meteorological Society Section: Journal of Climate.
- Rich Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, July 1997. ISSN 1573-0565. doi: 10.1023/A:1007379606734. URL <https://doi.org/10.1023/A:1007379606734>.

## Bibliography

---

- Jianxin Cheng, Jin Liu, Qiuming Kuang, Zhou Xu, Chenkai Shen, Wang Liu, and Kang Zhou. DeepDT: Generative Adversarial Network for High-Resolution Climate Prediction. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. ISSN 1558-0571. doi: 10.1109/LGRS.2020.3041760. Conference Name: IEEE Geoscience and Remote Sensing Letters.
- Jens H. Christensen, Fredrik Boberg, Ole B. Christensen, and Philippe Lucas-Picher. On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophysical Research Letters*, 35(20), 2008. ISSN 1944-8007. doi: 10.1029/2008GL035694. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2008GL035694>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2008GL035694>.
- Antje Claussnitzer. *Statistisch-Dynamische Analyse skalenabhängiger Niederschlagsprozesse: Vergleich zwischen Beobachtungen und Modell*. PhD thesis, Free University of Berlin, Berlin, 2010. URL <https://refubium.fu-berlin.de/handle/fub188/4202>. Accepted: 2018-06-07T17:43:51Z.
- Emmet Cleary, Alfredo Garbuno-Inigo, Shiwei Lan, Tapio Schneider, and Andrew M. Stuart. Calibrate, emulate, sample. *Journal of Computational Physics*, 424:109716, January 2021. ISSN 0021-9991. doi: 10.1016/j.jcp.2020.109716. URL <https://www.sciencedirect.com/science/article/pii/S0021999120304903>.
- Aiguo Dai. Global Precipitation and Thunderstorm Frequencies. Part I: Seasonal and Interannual Variations. *Journal of Climate*, 14(6):1092–1111, March 2001. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(2001)014<1092:GPATFP>2.0.CO;2. URL [https://journals.ametsoc.org/view/journals/clim/14/6/1520-0442\\_2001\\_014\\_1092\\_gpatfp\\_2.0.co\\_2.xml](https://journals.ametsoc.org/view/journals/clim/14/6/1520-0442_2001_014_1092_gpatfp_2.0.co_2.xml). Publisher: American Meteorological Society Section: Journal of Climate.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL <http://arxiv.org/abs/2010.11929>. arXiv:2010.11929 [cs].
- Markus Druke, Werner von Bloh, Stefan Petri, Boris Sakschewski, Sibyll Schaphoff, Matthias Forkel, Willem Huiskamp, Georg Feulner, and Kirsten Thonicke. CM2Mc-LPJmL v1.0: biophysical coupling of a process-based dynamic vegetation model with managed land to a general circulation model. *Geoscientific Model Development*, 14(6):4117–4141, July 2021. ISSN 1991-959X. doi: 10.5194/gmd-14-4117-2021. URL <https://gmd.copernicus.org/articles/14/4117/2021/>. Publisher: Copernicus GmbH.
- David Dunkerley. Intra-event intermittency of rainfall: an analysis of the metrics of rain and no-rain periods. *Hydrological Processes*, 29(15):3294–3305, 2015. ISSN 1099-1085. doi: 10.1002/hyp.10454. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.10454>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.10454>.
- J. P. Dunne, L. W. Horowitz, A. J. Adcroft, P. Ginoux, I. M. Held, J. G. John, J. P. Krasting, S. Malyshev, V. Naik, F. Paulot, E. Shevliakova, C. A. Stock, N. Zadeh, V. Balaji, C. Blanton, K. A. Dunne, C. Dupuis, J. Durachta, R. Dussin, P. P. G. Gauthier, S. M. Griffies, H. Guo, R. W. Hallberg, M. Harrison, J. He, W. Hurlin, C. McHugh, R. Menzel, P. C. D. Milly, S. Nikonov, D. J. Paynter, J. Ploshay, A. Radhakrishnan, K. Rand, B. G. Reichl, T. Robinson, D. M. Schwarzkopf, L. T. Sentman, S. Underwood, H. Vahlenkamp, M. Winton, A. T. Wittenberg, B. Wyman, Y. Zeng, and M. Zhao. The GFDL Earth System Model Version 4.1

- (GFDL-ESM 4.1): Overall Coupled Model Description and Simulation Characteristics. *Journal of Advances in Modeling Earth Systems*, 12(11):e2019MS002015, 2020. ISSN 1942-2466. doi: 10.1029/2019MS002015. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002015>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS002015>.
- Michel Déqué. Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global and Planetary Change*, 57(1-2):16–26, 2007. ISSN 09218181. doi: 10.1016/j.gloplacha.2006.11.030.
- Imme Ebert-Uphoff and Kyle Hilburn. Evaluation, Tuning, and Interpretation of Neural Networks for Working with Images in Meteorological Applications. *Bulletin of the American Meteorological Society*, 101(12):E2149–E2170, December 2020. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-20-0097.1. URL <https://journals.ametsoc.org/view/journals/bams/101/12/BAMS-D-20-0097.1.xml>. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.
- Jonathan M. Eden, Martin Widmann, David Grawe, and Sebastian Rast. Skill, Correction, and Downscaling of GCM-Simulated Precipitation. *Journal of Climate*, 25(11):3970–3984, June 2012. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI-D-11-00254.1. URL <https://journals.ametsoc.org/view/journals/clim/25/11/jcli-d-11-00254.1.xml>. Publisher: American Meteorological Society Section: Journal of Climate.
- European Centre for Medium-Range Weather Forecasts. The ECMWF ensemble prediction system, 2012. URL [https://www.ecmwf.int/sites/default/files/the\\_ECMWF\\_Ensemble\\_prediction\\_system.pdf](https://www.ecmwf.int/sites/default/files/the_ECMWF_Ensemble_prediction_system.pdf).
- European Centre for Medium-Range Weather Forecasts. PART III : Dynamics and Numerical Procedures. *IFS Documentation Cy46r1*, 3(June 2019):1–29, 2019. URL <https://www.ecmwf.int/node/19307>.
- Veronika Eyring, Peter M. Cox, Gregory M. Flato, Peter J. Gleckler, Gab Abramowitz, Peter Caldwell, William D. Collins, Bettina K. Gier, Alex D. Hall, Forrest M. Hoffman, George C. Hurtt, Alexandra Jahn, Chris D. Jones, Stephen A. Klein, John P. Krasting, Lester Kwiatkowski, Ruth Lorenz, Eric Maloney, Gerald A. Meehl, Angeline G. Pendergrass, Robert Pincus, Alex C. Ruane, Joellen L. Russell, Benjamin M. Sanderson, Benjamin D. Santer, Steven C. Sherwood, Isla R. Simpson, Ronald J. Stouffer, and Mark S. Williamson. Taking climate model evaluation to the next level. *Nature Climate Change*, 9(2):102–110, 2019. ISSN 17586798. doi: 10.1038/s41558-018-0355-y. URL <http://dx.doi.org/10.1038/s41558-018-0355-y>. Publisher: Springer US.
- E. M. Fischer and R. Knutti. Observed heavy precipitation increase confirms theory and early models. *Nature Climate Change*, 6(11):986–991, November 2016. ISSN 1758-6798. doi: 10.1038/nclimate3110. URL <https://www.nature.com/articles/nclimate3110>. Number: 11 Publisher: Nature Publishing Group.
- Gabriele Franch, Valerio Maggio, Luca Coviello, Marta Pendesini, Giuseppe Jurman, and Cesare Furlanello. TAASRAD19, a high-resolution weather radar reflectivity dataset for precipitation nowcasting. *Scientific Data*, 7(1):234, July 2020a. ISSN 2052-4463. doi: 10.1038/s41597-020-0574-8. URL <https://www.nature.com/articles/s41597-020-0574-8>. Number: 1 Publisher: Nature Publishing Group.
- Gabriele Franch, Daniele Nerini, Marta Pendesini, Luca Coviello, Giuseppe Jurman, and Cesare Furlanello. Precipitation Nowcasting with Orographic Enhanced Stacked Generalization: Improving Deep Learning Predictions on Extreme Events. *Atmosphere*, 11(3):267, March 2020b. ISSN 2073-4433. doi: 10.3390/atmos11030267. URL <https://www.mdpi.com/2073-4433/11/3/267>. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

## Bibliography

---

- Bastien François, Soulivanh Thao, and Mathieu Vrac. *Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks*. Springer Berlin Heidelberg, 2021. ISBN 0-12-345678-9. doi: 10.1007/s00382-021-05869-8. URL <https://doi.org/10.1007/s00382-021-05869-8>. Publication Title: Climate Dynamics Issue: 0123456789 ISSN: 14320894.
- Christoph Frei and Christoph Schär. Detection Probability of Trends in Rare Events: Theory and Application to Heavy Precipitation in the Alpine Region. *Journal of Climate*, 14(7):1568–1584, April 2001. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(2001)014<1568:DPOTIR>2.0.CO;2. URL [https://journals.ametsoc.org/view/journals/clim/14/7/1520-0442\\_2001\\_014\\_1568\\_dpotir\\_2.0.co\\_2.xml](https://journals.ametsoc.org/view/journals/clim/14/7/1520-0442_2001_014_1568_dpotir_2.0.co_2.xml). Publisher: American Meteorological Society Section: Journal of Climate.
- James Fulton and Ben Clarke. Towards debiasing climate simulations using unsupervised image-to-image translation networks. In *Climate Change AI*. Climate Change AI, December 2021. URL <https://www.climatechange.ai/papers/neurips2021/7>.
- Alan J. Geer, Katrin Lonitz, Peter Weston, Masahiro Kazumori, Kozo Okamoto, Yanqiu Zhu, Emily Huichun Liu, Andrew Collard, William Bell, Stefano Migliorini, Philippe Chambon, Nadia Fourrié, Min-Jeong Kim, Christina Köpken-Watts, and Christoph Schraff. All-sky satellite data assimilation at operational weather forecasting centres. *Quarterly Journal of the Royal Meteorological Society*, 144(713):1191–1217, 2018. ISSN 1477-870X. doi: 10.1002/qj.3202. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3202>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3202>.
- P. Gentine, M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis. Could Machine Learning Break the Convection Parameterization Deadlock? *Geophysical Research Letters*, 45(11):5742–5751, June 2018. ISSN 19448007. doi: 10.1029/2018GL078202. URL <https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2018GL078202>. Publisher: Blackwell Publishing Ltd.
- Harry R. Glahn and Dale A. Lowry. The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology and Climatology*, 11(8):1203–1211, December 1972. ISSN 1520-0450. doi: 10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2. URL [https://journals.ametsoc.org/view/journals/apme/11/8/1520-0450\\_1972\\_011\\_1203\\_tuomos\\_2\\_0\\_co\\_2.xml](https://journals.ametsoc.org/view/journals/apme/11/8/1520-0450_1972_011_1203_tuomos_2_0_co_2.xml). Publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology.
- Tilmann Gneiting and Roopesh Ranjan. Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*, 29(3):411–422, July 2011. ISSN 0735-0015. doi: 10.1198/jbes.2010.08110. URL <https://doi.org/10.1198/jbes.2010.08110>. Publisher: Taylor & Francis \_eprint: <https://doi.org/10.1198/jbes.2010.08110>.
- Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005. ISSN 00270644. doi: 10.1175/MWR2904.1.
- Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2007.00587.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00587.x>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2007.00587.x>.
- Aaron Gokaslan, Vivek Ramanujan, Daniel Ritchie, Kwang In Kim, and James Tompkin. Improving Shape Deformation in Unsupervised Image-to-Image Translation.

- In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 649–665, 2018. URL [https://openaccess.thecvf.com/content\\_ECCV\\_2018/html/Aaron\\_Gokaslan\\_Improving\\_Shape\\_Deformation\\_ECCV\\_2018\\_paper.html](https://openaccess.thecvf.com/content_ECCV_2018/html/Aaron_Gokaslan_Improving_Shape_Deformation_ECCV_2018_paper.html).
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, November 2016. ISBN 978-0-262-33737-3. Google-Books-ID: omivDQAAQBAJ.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, October 2020. ISSN 0001-0782, 1557-7317. doi: 10.1145/3422622. URL <https://dl.acm.org/doi/10.1145/3422622>.
- Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian Neural Networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/26cd8ecadce0d4efd6cc8a8725cbd1f8-Paper.pdf>.
- Peter Grönquist, Chengyuan Yao, Tal Ben-Nun, Nikoli Dryden, Peter Dueben, Shigang Li, Torsten Hoefler, Eth Zürich, Chengyuan Yao, Tal Ben-Nun, Nikoli Dryden, Peter Dueben, Shigang Li, and Torsten Hoefler. Deep Learning for Post-Processing Ensemble Weather Forecastse, May 2020. URL <https://cds.climate.copernicus.eu/>.
- L. Gudmundsson, J. B. Bremnes, J. E. Haugen, and T. Engen-Skaugen. Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations &ndash; a comparison of methods. *Hydrology and Earth System Sciences*, 16(9):3383–3390, September 2012. ISSN 1027-5606. doi: 10.5194/hess-16-3383-2012. URL <https://hess.copernicus.org/articles/16/3383/2012/>. Publisher: Copernicus GmbH.
- Selma B. Guerreiro, Hayley J. Fowler, Renaud Barbero, Seth Westra, Geert Lenderink, Stephen Blenkinsop, Elizabeth Lewis, and Xiao Feng Li. Detection of continental-scale intensification of hourly rainfall extremes. *Nature Climate Change*, 8(9):803–807, September 2018. ISSN 17586798. doi: 10.1038/s41558-018-0245-3. URL <https://doi.org/10.1038/s41558-018-0245-3>. Publisher: Nature Publishing Group.
- Reindert J. Haarsma, Malcolm J. Roberts, Pier Luigi Vidale, Catherine A. Senior, Alessio Bellucci, Qing Bao, Ping Chang, Susanna Corti, Neven S. Fućkar, Virginie Guemas, Jost von Hardenberg, Wilco Hazeleger, Chihiro Kodama, Torben Koenigk, L. Ruby Leung, Jian Lu, Jing-Jia Luo, Jiafu Mao, Matthew S. Mizielinski, Ryo Mizuta, Paulo Nobre, Masaki Satoh, Enrico Scoccimarro, Tido Semmler, Justin Small, and Jin-Song von Storch. High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6. *Geoscientific Model Development*, 9(11):4185–4208, November 2016. ISSN 1991-9603. doi: 10.5194/gmd-9-4185-2016. URL <https://gmd.copernicus.org/articles/9/4185/2016/>.
- Alex Hall, Peter Cox, Chris Huntingford, and Stephen Klein. Progressing emergent constraints on future climate change. *Nature Climate Change*, 9(4):269–278, 2019. ISSN 17586798. doi: 10.1038/s41558-019-0436-6. URL <http://dx.doi.org/10.1038/s41558-019-0436-6>. Publisher: Springer US.
- Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. Deep learning for multi-year ENSO forecasts. *Nature*, 573(7775):568–572, September 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1559-7. URL <http://www.nature.com/articles/s41586-019-1559-7>.

## Bibliography

---

- Lucy Harris, Andrew T. T. McRae, Matthew Chantry, Peter D. Dueben, and Tim N. Palmer. A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts. *Journal of Advances in Modeling Earth Systems*, 14(10), October 2022. ISSN 1942-2466, 1942-2466. doi: 10.1029/2022MS003120. URL <http://arxiv.org/abs/2204.02028>. arXiv:2204.02028 [physics, stat].
- R Giles Harrison, Keri A Nicoll, Maarten H P Ambaum, Graeme J Marlton, Karen L Aplin, and Michael Lockwood. Precipitation Modification by Ionization. *Physical Review Letters*, 124(19):198701, 2020. ISSN 1079-7114. doi: 10.1103/PhysRevLett.124.198701. URL <https://doi.org/10.1103/PhysRevLett.124.198701>. Publisher: American Physical Society.
- Birgit Hassler and Axel Lauer. Comparison of Reanalysis and Observational Precipitation Datasets Including ERA5 and WFDE5. *Atmosphere*, 12(11):1462, November 2021. ISSN 2073-4433. doi: 10.3390/atmos12111462. URL <https://www.mdpi.com/2073-4433/12/11/1462>. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778. IEEE Computer Society, December 2016. ISBN 978-1-4673-8850-4. doi: 10.1109/CVPR.2016.90. arXiv: 1512.03385 ISSN: 10636919.
- Isaac M. Held and Brian J. Soden. Robust Responses of the Hydrological Cycle to Global Warming. *Journal of Climate*, 19(21):5686–5699, November 2006. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI3990.1. URL <https://journals.ametsoc.org/view/journals/clim/19/21/jcli3990.1.xml>. Publisher: American Meteorological Society Section: Journal of Climate.
- Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, July 2020. ISSN 1477870X. doi: 10.1002/qj.3803. URL <https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.3803>. Publisher: John Wiley and Sons Ltd.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html>.
- Brian Hoskins. The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Quarterly Journal of the Royal Meteorological Society*, 139(672):573–584, 2013. ISSN 1477-870X. doi: 10.1002/qj.1991. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.1991>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.1991>.
- Arthur Y. Hou, Ramesh K. Kakar, Steven Neeck, Ardeshir A. Azarbarzin, Christian D. Kummerow, Masahiro Kojima, Riko Oki, Kenji Nakamura, and Toshio Iguchi. The Global Precipitation Measurement Mission. *Bulletin of the American Meteorological Society*, 95(5):701–722, May 2014. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-13-00164.1. URL <https://journals.ametsoc.org/view/journals/bams/95/5/bams-d-13-00164.1.xml>. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

- Frédéric Hourdin, Thorsten Mauritsen, Andrew Gettelman, Jean Christophe Golaz, Venkatramani Balaji, Qingyun Duan, Doris Folini, Duoying Ji, Daniel Klocke, Yun Qian, Florian Rauser, Catherine Rio, Lorenzo Tomassini, Masahiro Watanabe, and Daniel Williamson. The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, 98(3): 589–602, 2017. ISSN 00030007. doi: 10.1175/BAMS-D-15-00135.1.
- Robert A. Houze Jr. Orographic effects on precipitating clouds. *Reviews of Geophysics*, 50(1), 2012. ISSN 1944-9208. doi: 10.1029/2011RG000365. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2011RG000365>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2011RG000365>.
- Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal Unsupervised Image-to-image Translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. URL [https://openaccess.thecvf.com/content\\_ECCV\\_2018/html/Xun\\_Huang\\_Multimodal\\_Unsupervised\\_Image-to-image\\_ECCV\\_2018\\_paper.html](https://openaccess.thecvf.com/content_ECCV_2018/html/Xun_Huang_Multimodal_Unsupervised_Image-to-image_ECCV_2018_paper.html).
- George J. Huffman, David T. Bolvin, Eric J. Nelkin, David B. Wolff, Robert F. Adler, Guojun Gu, Yang Hong, Kenneth P. Bowman, and Erich F. Stocker. The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. *Journal of Hydrometeorology*, 8(1):38–55, February 2007. ISSN 1525-7541, 1525-755X. doi: 10.1175/JHM560.1. URL [https://journals.ametsoc.org/view/journals/hydr/8/1/jhm560\\_1.xml](https://journals.ametsoc.org/view/journals/hydr/8/1/jhm560_1.xml). Publisher: American Meteorological Society Section: Journal of Hydrometeorology.
- Chris Huntingford, Elizabeth S Jeffers, Michael B Bonsall, Hannah M Christensen, Thomas Lees, and Hui Yang. Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, 14(12):124007, November 2019. ISSN 1748-9326. doi: 10.1088/1748-9326/ab4e55. URL <https://iopscience.iop.org/article/10.1088/1748-9326/ab4e55>.
- Dion Häfner, Roman Nuterman, and Markus Jochum. Fast, Cheap, and Turbulent—Global Ocean Modeling With GPU Acceleration in Python. *Journal of Advances in Modeling Earth Systems*, 13(12):e2021MS002717, 2021. ISSN 1942-2466. doi: 10.1029/2021MS002717. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002717>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021MS002717>.
- IPCC. Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Technical report, Cambridge University Press, 2021. URL <https://www.ipcc.ch/report/sixth-assessment-report-working-group-i/>.
- Christopher Irrgang, Niklas Boers, Maike Sonnewald, Elizabeth A. Barnes, Christopher Kadow, Joanna Staneva, and Jan Saynisch-Wagner. Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. *Nature Machine Intelligence*, 3(8):667–674, August 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00374-3. URL <https://www.nature.com/articles/s42256-021-00374-3>. Number: 8 Publisher: Nature Publishing Group.
- Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017. ISSN 10414347. doi: 10.1109/TKDE.2017.2720168. arXiv: 1612.08544.
- Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. URL [https://openaccess.thecvf.com/content\\_cvpr\\_2019/html/Karras\\_A\\_Style-Based\\_Generator\\_Architecture\\_for\\_Generative\\_Adversarial\\_Networks.html](https://openaccess.thecvf.com/content_cvpr_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks.html).



## Bibliography

---

[com/content\\_CVPR\\_2019/html/Karras\\_A\\_Style-Based\\_Generator\\_Architecture\\_for\\_Generative\\_Adversarial\\_Networks\\_CVPR\\_2019\\_paper.html](https://arxiv.org/abs/2202.07575).

Ryan Keisler. Forecasting Global Weather with Graph Neural Networks, February 2022. URL <http://arxiv.org/abs/2202.07575>. arXiv:2202.07575 [physics].

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, 2015. arXiv: 1412.6980.

Reto Knutti, Jan Sedláček, Benjamin M. Sanderson, Ruth Lorenz, Erich M. Fischer, and Veronika Eyring. A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, 44(4):1909–1918, 2017. ISSN 19448007. doi: 10.1002/2016GL072012.

Dmitrii Kochkov, Jamie A. Smith, Ayya Alieva, Qing Wang, Michael P. Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, May 2021. doi: 10.1073/pnas.2101784118. URL <https://www.pnas.org/doi/10.1073/pnas.2101784118>. Publisher: Proceedings of the National Academy of Sciences.

Maximilian Kotz, Anders Levermann, and Leonie Wenz. The effect of rainfall changes on economic production. *Nature*, 601(7892):223–227, January 2022. ISSN 1476-4687. doi: 10.1038/s41586-021-04283-8. URL <https://www.nature.com/articles/s41586-021-04283-8>. Number: 7892 Publisher: Nature Publishing Group.

Demetris Koutsoyiannis. Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation. *Hydrological Sciences Journal*, 49(4):575–590, August 2004a. ISSN 02626667. doi: 10.1623/hysj.49.4.575.54430. URL <https://www.tandfonline.com/doi/abs/10.1623/hysj.49.4.575.54430>. Publisher: IAHS Press.

Demetris Koutsoyiannis. Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records. *Hydrological Sciences Journal*, 49(4):591–610, August 2004b. ISSN 02626667. doi: 10.1623/hysj.49.4.591.54424. URL <https://www.tandfonline.com/doi/abs/10.1623/hysj.49.4.591.54424>. Publisher: IAHS Press.

Basil Kraft, Martin Jung, Marco Körner, Sujan Koirala, and Markus Reichstein. Towards hybrid modeling of the global hydrological cycle. *Hydrology and Earth System Sciences*, 26(6): 1579–1614, March 2022. ISSN 1027-5606. doi: 10.5194/hess-26-1579-2022. URL <https://hess.copernicus.org/articles/26/1579/2022/>. Publisher: Copernicus GmbH.

John P Krasting, Jasmin G John, Chris Blanton, Colleen McHugh, Serguei Nikonov, Aparna Radhakrishnan, Kristopher Rand, Niki T Zadeh, V Balaji, Jeff Durachta, Christopher Dupuis, Raymond Menzel, Thomas Robinson, Seth Underwood, Hans Vahlenkamp, Krista A Dunne, Paul P G Gauthier, Paul Ginoux, Stephen M Griffies, Robert Hallberg, Matthew Harrison, William Hurlin, Sergey Malyshev, Vaishali Naik, Fabien Paulot, David J Paynter, Jeffrey Ploshay, Brandon G Reichl, Daniel M Schwarzkopf, Charles J Seman, Levi Silvers, Bruce Wyman, Yujin Zeng, Alistair Adcroft, John P Dunne, Raphael Dussin, Huan Guo, Jian He, Isaac M Held, Larry W Horowitz, Pu Lin, P C D Milly, Elena Shevliakova, Charles Stock, Michael Winton, Andrew T Wittenberg, Yuanyu Xie, and Ming Zhao. NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 CMIP, 2018. URL <https://doi.org/10.22033/ESGF/CMIP6.1407>.

Paul A. Kucera, Elizabeth E. Ebert, F. Joseph Turk, Vincenzo Levizzani, Dalia Kirschbaum, Francisco J. Tapiador, Alexander Loew, and M. Borsche. Precipitation from Space: Advancing Earth System Science. *Bulletin of the American Meteorological Society*, 94(3):365–375, March 2013. doi: 10.1175/BAMS-D-11-00171.1. URL <https://journals.ametsoc.org/view/journals/bams/94/3/bams-d-11-00171.1.xml>. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

- Christian Kummerow, William Barnes, Toshiaki Kozu, James Shiue, and Joanne Simpson. The Tropical Rainfall Measuring Mission (TRMM) Sensor Package. *Journal of Atmospheric and Oceanic Technology*, 15(3):809–817, June 1998. ISSN 0739-0572, 1520-0426. doi: 10.1175/1520-0426(1998)015<0809:TTRMMT>2.0.CO;2. URL [https://journals.ametsoc.org/view/journals/atot/15/3/1520-0426\\_1998\\_015\\_0809\\_ttrmmt\\_2\\_0\\_co\\_2.xml](https://journals.ametsoc.org/view/journals/atot/15/3/1520-0426_1998_015_0809_ttrmmt_2_0_co_2.xml). Publisher: American Meteorological Society Section: Journal of Atmospheric and Oceanic Technology.
- Stefan Lange. Trend-preserving bias adjustment and statistical downscaling with ISIMIP3BASD (v1.0). *Geoscientific Model Development*, 12(7):3055–3070, 2019. ISSN 19919603. doi: 10.5194/gmd-12-3055-2019.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL <https://www.nature.com/articles/nature14539>. Number: 7553 Publisher: Nature Publishing Group.
- Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient BackProp. In Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade: Second Edition*, Lecture Notes in Computer Science, pages 9–48. Springer, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8\_3. URL [https://doi.org/10.1007/978-3-642-35289-8\\_3](https://doi.org/10.1007/978-3-642-35289-8_3).
- Jascha Lehmann, Dim Coumou, and Katja Frieler. Increased record-breaking precipitation events under global warming. *Climatic Change*, 132(4):501–515, October 2015. ISSN 1573-1480. doi: 10.1007/s10584-015-1434-y. URL <https://doi.org/10.1007/s10584-015-1434-y>.
- Sebastian Lerch and Thordis L. Thorarinsdottir. Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A: Dynamic Meteorology and Oceanography*, 65(1):21206, December 2013. ISSN null. doi: 10.3402/tellusa.v65i0.21206. URL <https://doi.org/10.3402/tellusa.v65i0.21206>. Publisher: Taylor & Francis \_eprint: <https://doi.org/10.3402/tellusa.v65i0.21206>.
- Sebastian Lerch, Thordis L. Thorarinsdottir, Francesco Ravazzolo, and Tilmann Gneiting. Forecaster’s dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32(1):106–127, 2017. ISSN 08834237. doi: 10.1214/16-STS588. arXiv: 1512.09244.
- Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Aizzadenesheli, and Anima Anandkumar. Physics-Informed Neural Operator for Learning Partial Differential Equations, November 2022. URL <http://arxiv.org/abs/2111.03794>. arXiv:2111.03794 [cs, math].
- Xiaodan Liang, Hao Zhang, Liang Lin, and Eric Xing. Generative Semantic Manipulation with Mask-Contrasting GAN. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 558–573, 2018. URL [https://openaccess.thecvf.com/content\\_ECCV\\_2018/html/Liang\\_Generative\\_Semantic\\_Manipulation\\_ECCV\\_2018\\_paper.html](https://openaccess.thecvf.com/content_ECCV_2018/html/Liang_Generative_Semantic_Manipulation_ECCV_2018_paper.html).
- Julia Ling, Reese Jones, and Jeremy Templeton. Machine learning strategies for systems with invariance properties. *Journal of Computational Physics*, 318:22–35, August 2016. ISSN 0021-9991. doi: 10.1016/j.jcp.2016.05.003. URL <https://www.sciencedirect.com/science/article/pii/S0021999116301309>.
- Zachary C. Lipton. The Mythos of Model Interpretability, March 2017. URL <http://arxiv.org/abs/1606.03490>. arXiv:1606.03490 [cs, stat].
- Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised Image-to-Image Translation Networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/dc6a6489640ca02b0d42dabeb8e46bb7-Abstract.html>.

## Bibliography

---

- Ignacio Lopez-Gomez, Amy McGovern, Shreya Agrawal, and Jason Hickey. Global Extreme Heat Forecasting Using Neural Weather Models, November 2022. URL <http://arxiv.org/abs/2205.10972>. arXiv:2205.10972 [physics].
- Edward N. Lorenz. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, March 1963. ISSN 0022-4928, 1520-0469. doi: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2. URL [https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469\\_1963\\_020\\_0130\\_dnf\\_2\\_0\\_co\\_2.xml](https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml). Publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences.
- Edward N. Lorenz. On the possible reasons for long-period fluctuations of the general circulation. In *Proc. WMO-IUGG Symp. on Research and Development Aspects of Long-Range Forecasting.*, 1965.
- S. Lovejoy, D. Schertzer, and A. A. Tsonis. Functional Box-Counting and Multiple Elliptical Dimensions in Rain. *Science*, 235(4792):1036–1038, February 1987. doi: 10.1126/science.235.4792.1036. URL <https://www.science.org/doi/abs/10.1126/science.235.4792.1036>. Publisher: American Association for the Advancement of Science.
- Benoit Mandelbrot. How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension. *Science*, 156(3775):636–638, May 1967. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.156.3775.636. URL <https://www.science.org/doi/10.1126/science.156.3775.636>.
- D. Maraun. Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. *Geophysical Research Letters*, 39(6), 2012. ISSN 1944-8007. doi: 10.1029/2012GL051210. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2012GL051210>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2012GL051210>.
- Jeremy McGibbon, Noah D. Brenowitz, Mark Cheeseman, Spencer K. Clark, Johann P. S. Dahm, Eddie C. Davis, Oliver D. Elbert, Rhea C. George, Lucas M. Harris, Brian Henn, Anna Kwa, W. Andre Perkins, Oliver Watt-Meyer, Tobias F. Wicky, Christopher S. Bretherton, and Oliver Fuhrer. fv3gfs-wrapper: a Python wrapper of the FV3GFS atmospheric model. *Geoscientific Model Development*, 14(7):4401–4409, July 2021. ISSN 1991-9603. doi: 10.5194/gmd-14-4401-2021. URL <https://gmd.copernicus.org/articles/14/4401/2021/>.
- Amy McGovern, Ryan Lagerquist, David John Gagne, G. Eli Jergensen, Kimberly L. Elmore, Cameron R. Homeyer, and Travis Smith. Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bulletin of the American Meteorological Society*, 100(11):2175–2199, November 2019. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-18-0195.1. URL <https://journals.ametsoc.org/view/journals/bams/100/11/bams-d-18-0195.1.xml>. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.
- Michael J. McPhaden, Stephen E. Zebiak, and Michael H. Glantz. ENSO as an Integrating Concept in Earth Science. *Science*, 314(5806):1740–1745, December 2006. doi: 10.1126/science.1132588. URL <https://www.science.org/doi/full/10.1126/science.1132588>. Publisher: American Association for the Advancement of Science.
- L. V. Meisel, Mark Johnson, and P. J. Cote. Box-counting multifractal analysis. *Physical Review A*, 45(10):6989–6996, May 1992. doi: 10.1103/PhysRevA.45.6989. URL <https://link.aps.org/doi/10.1103/PhysRevA.45.6989>. Publisher: American Physical Society.
- S. Michaelides, V. Levizzani, E. Anagnostou, P. Bauer, T. Kasparis, and J.E. Lane. Precipitation: Measurement, remote sensing, climatology and modeling. *Atmospheric Research*, 94(4): 512–533, December 2009. ISSN 01698095. doi: 10.1016/j.atmosres.2009.08.017. URL <https://linkinghub.elsevier.com/retrieve/pii/S0169809509002488>.

- Kingtse C. Mo and R. W. Higgins. Tropical Convection and Precipitation Regimes in the Western United States. *Journal of Climate*, 11(9):2404–2423, September 1998. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(1998)011<2404:TCAPRI>2.0.CO;2. URL [https://journals.ametsoc.org/view/journals/clim/11/9/1520-0442\\_1998\\_011\\_2404\\_tcapri\\_2.0.co\\_2.xml](https://journals.ametsoc.org/view/journals/clim/11/9/1520-0442_1998_011_2404_tcapri_2.0.co_2.xml). Publisher: American Meteorological Society Section: Journal of Climate.
- Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. In Irena Koprinska, Michael Kamp, Annalisa Appice, Corrado Loglisci, Luiza Antonie, Albrecht Zimmermann, Riccardo Guidotti, Özlem Özgöbek, Rita P. Ribeiro, Ricard Gavaldà, João Gama, Linara Adilova, Yamuna Krishnamurthy, Pedro M. Ferreira, Donato Malerba, Ibéria Medeiros, Michelangelo Ceci, Giuseppe Manco, Elio Masciari, Zbigniew W. Ras, Peter Christen, Eirini Ntoutsis, Erich Schubert, Arthur Zimek, Anna Monreale, Przemyslaw Biecek, Salvatore Rinzivillo, Benjamin Kille, Andreas Lommatzsch, and Jon Atle Gulla, editors, *ECML PKDD 2020 Workshops, Communications in Computer and Information Science*, pages 417–431, Cham, 2020. Springer International Publishing. ISBN 978-3-030-65965-3. doi: 10.1007/978-3-030-65965-3\_28.
- Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, February 2018. ISSN 1051-2004. doi: 10.1016/j.dsp.2017.10.011. URL <https://www.sciencedirect.com/science/article/pii/S1051200417302385>.
- Ionel M. Navon. Data Assimilation for Numerical Weather Prediction: A Review. In Seon K. Park and Liang Xu, editors, *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications*, pages 21–65. Springer, Berlin, Heidelberg, 2009. ISBN 978-3-540-71056-1. doi: 10.1007/978-3-540-71056-1\_2. URL [https://doi.org/10.1007/978-3-540-71056-1\\_2](https://doi.org/10.1007/978-3-540-71056-1_2).
- Paul A. O’Gorman and John G. Dwyer. Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events. *Journal of Advances in Modeling Earth Systems*, 10(10):2548–2563, October 2018. ISSN 1942-2466, 1942-2466. doi: 10.1029/2018MS001351. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001351>.
- Tim Palmer. The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, 145(S1):12–24, September 2019. ISSN 0035-9009. doi: 10.1002/qj.3383. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3383>.
- Tim Palmer. A Vision for Numerical Weather Prediction in 2030, April 2022. URL <http://arxiv.org/abs/2007.04830>. arXiv:2007.04830 [physics].
- Tim Palmer and Bjorn Stevens. The scientific challenge of understanding and estimating climate change. *Proceedings of the National Academy of Sciences of the United States of America*, 116(49):34390–34395, 2019. ISSN 10916490. doi: 10.1073/pnas.1906691116.
- T.N. Palmer. Extended-Range Atmospheric Prediction and the Lorenz Model. *Bulletin of the American Meteorological Society*, 74(1):49–65, 1993. ISSN 0003-0007. URL <https://www.jstor.org/stable/26230418>. Publisher: American Meteorological Society.
- Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, February 2022. URL <http://arxiv.org/abs/2202.11214>. arXiv:2202.11214 [physics].

## Bibliography

---

- Nathanaël Perraudin, Michaël Defferrard, Tomasz Kacprzak, and Raphael Sgier. DeepSphere: Efficient spherical Convolutional Neural Network with HEALPix sampling for cosmological applications. *arXiv:1810.12186 [astro-ph]*, March 2019. URL <http://arxiv.org/abs/1810.12186>. arXiv: 1810.12186.
- S. Pfahl, P. A. O’Gorman, and E. M. Fischer. Understanding the regional pattern of projected future changes in extreme precipitation. *Nature Climate Change*, 7(6):423–427, June 2017. ISSN 1758-6798. doi: 10.1038/nclimate3287. URL <https://www.nature.com/articles/nclimate3287>. Number: 6 Publisher: Nature Publishing Group.
- Lutz Prechelt. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767, June 1998. ISSN 0893-6080. doi: 10.1016/S0893-6080(98)00010-0. URL <https://www.sciencedirect.com/science/article/pii/S0893608098000100>.
- Ilan Price and Stephan Rasp. Increasing the accuracy and resolution of precipitation forecasts using deep generative models, March 2022. URL <http://arxiv.org/abs/2203.12297>. arXiv:2203.12297 [cs, stat].
- Di Qi and Andrew J. Majda. Using machine learning to predict extreme events in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, 117(1):52–59, January 2020. ISSN 10916490. doi: 10.1073/pnas.1917285117. URL [www.pnas.org/cgi/doi/10.1073/pnas.1917285117](http://www.pnas.org/cgi/doi/10.1073/pnas.1917285117). Publisher: National Academy of Sciences.
- Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, January 2016. URL <http://arxiv.org/abs/1511.06434>. arXiv:1511.06434 [cs].
- M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, February 2019. ISSN 00219991. doi: 10.1016/j.jcp.2018.10.045. URL <https://linkinghub.elsevier.com/retrieve/pii/S0021999118307125>.
- Stephan Rasp. *Statistical methods and machine learning in weather and climate modeling*. Text.PhDThesis, Ludwig-Maximilians-Universität München, March 2019. URL <https://edoc.ub.uni-muenchen.de/23867/>. ISSN: 1923-8673.
- Stephan Rasp and Nils Thuerey. Data-driven medium-range weather prediction with a Resnet pretrained on climate simulations: A new model for WeatherBench. *Journal of Advances in Modeling Earth Systems*, February 2021. ISSN 1942-2466. doi: 10.1029/2020ms002405. arXiv: 2008.08626 Publisher: American Geophysical Union (AGU).
- Stephan Rasp, Michael S. Pritchard, and Pierre Gentine. Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39):9684–9689, September 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1810286115. URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1810286115>.
- Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021. ISSN 14764687. doi: 10.1038/s41586-021-03854-z. URL <http://arxiv.org/abs/2104.00954>. arXiv: 2104.00954 Publisher: Springer US.
- Deep Ray, Orazio Pinti, and Assad A. Oberai. Deep Learning and Computational Physics (Lecture Notes), January 2023. URL <http://arxiv.org/abs/2301.00942>. arXiv:2301.00942 [math-ph].

- Abdul Rehman, Yang Gao, Jiheng Wang, and Zhou Wang. Image classification based on complex wavelet structural similarity. *Signal Processing: Image Communication*, 28(8):984–992, 2013. ISSN 09235965. doi: 10.1016/j.image.2012.07.004. URL <http://dx.doi.org/10.1016/j.image.2012.07.004>. Publisher: Elsevier.
- Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204, February 2019. ISSN 14764687. doi: 10.1038/s41586-019-0912-1. URL <http://www.nature.com/articles/s41586-019-0912-1>.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9351, pages 234–241. Springer Verlag, May 2015. ISBN 978-3-319-24573-7. doi: 10.1007/978-3-319-24574-4\_28. URL <http://lmb.informatik.uni-freiburg.de/>. arXiv: 1505.04597 ISSN: 16113349.
- Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, 8:42200–42216, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2976199. Conference Name: IEEE Access.
- F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1959. ISSN 1939-1471. doi: 10.1037/h0042519. URL <https://psycnet.apa.org/fulltext/1959-09865-001.pdf>. Publisher: US: American Psychological Association.
- Leon D. Rotstajn and Ulrike Lohmann. Tropical Rainfall Trends and the Indirect Aerosol Effect. *Journal of Climate*, 15(15):2103–2116, August 2002. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(2002)015<2103:TRTATI>2.0.CO;2. URL [https://journals.ametsoc.org/view/journals/clim/15/15/1520-0442\\_2002\\_015\\_2103\\_trtati\\_2.0.co\\_2.xml](https://journals.ametsoc.org/view/journals/clim/15/15/1520-0442_2002_015_2103_trtati_2.0.co_2.xml). Publisher: American Meteorological Society Section: Journal of Climate.
- David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. ISSN 1476-4687. doi: 10.1038/323533a0. URL <https://www.nature.com/articles/323533a0>. Number: 6088 Publisher: Nature Publishing Group.
- Anamitra Saha and Sai Ravela. Downscaling Extreme Rainfall Using Physical-Statistical Generative Adversarial Learning, December 2022. URL <http://arxiv.org/abs/2212.01446>. arXiv:2212.01446 [physics, stat].
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved Techniques for Training GANs. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL <https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf>.
- Mehul P Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K Markey. Complex Wavelet Structural Similarity: A New Image Similarity Index. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 18(11), 2009. doi: 10.1109/TIP.2009.2025923. URL <http://ieeexplore.ieee.org>.
- Clayton Sanford, Anna Kwa, Oliver Watt-Meyer, Spencer Clark, Noah Brenowitz, Jeremy McGibbon, and Christopher Bretherton. Improving the predictions of ML-corrected climate models with novelty detection, November 2022. URL <http://arxiv.org/abs/2211.13354>. arXiv:2211.13354 [physics].

## Bibliography

---

- Hiroshi Sasaki, Chris G. Willcocks, and Toby P. Breckon. UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models, April 2021. URL <http://arxiv.org/abs/2104.05358>. arXiv:2104.05358 [cs, eess].
- Sebastian Scher and Gabriele Messori. Spherical convolution and other forms of informed machine learning for deep neural network based weather forecasts, January 2021. URL <http://arxiv.org/abs/2008.13524>. arXiv:2008.13524 [physics].
- Daniel Schertzer and Shaun Lovejoy. Physical Modeling and Analysis of Rain and Clouds by Anisotropic Scaling Multiplicative Processes. *Journal of Geophysical Research*, 92:9693 – 9714, 1987.
- Marc Schleiss. How intermittency affects the rate at which rainfall extremes respond to changes in temperature. *Earth System Dynamics*, 9(3):955–968, July 2018. ISSN 2190-4987. doi: 10.5194/esd-9-955-2018. URL <https://esd.copernicus.org/articles/9/955/2018/>.
- Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61: 85–117, January 2015. ISSN 0893-6080. doi: 10.1016/j.neunet.2014.09.003. URL <https://www.sciencedirect.com/science/article/pii/S0893608014002135>.
- Tapio Schneider, Shiwei Lan, Andrew Stuart, and João Teixeira. Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations: EARTH SYSTEM MODELING 2.0. *Geophysical Research Letters*, 44(24): 12,396–12,417, December 2017a. ISSN 00948276. doi: 10.1002/2017GL076101. URL <http://doi.wiley.com/10.1002/2017GL076101>.
- Tapio Schneider, João Teixeira, Christopher S. Bretherton, Florent Brient, Kyle G. Pressel, Christoph Schär, and A. Pier Siebesma. Climate goals and computing the future of clouds. *Nature Climate Change*, 7(1):3–5, January 2017b. ISSN 1758-6798. doi: 10.1038/nclimate3190. URL <https://www.nature.com/articles/nclimate3190>. Number: 1 Publisher: Nature Publishing Group.
- M. G. Schultz, C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadler. Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 2021. ISSN 1364503X. doi: 10.1098/rsta.2020.0097. ISBN: 000000177702.
- Benedikt Schulz and Sebastian Lerch. Machine Learning Methods for Postprocessing Ensemble Forecasts of Wind Gusts: A Systematic Comparison. *Monthly Weather Review*, 150(1):235–257, January 2022. ISSN 1520-0493, 0027-0644. doi: 10.1175/MWR-D-21-0150.1. URL <https://journals.ametsoc.org/view/journals/mwre/150/1/MWR-D-21-0150.1.xml>. Publisher: American Meteorological Society Section: Monthly Weather Review.
- Francesco Serinaldi and Chris G. Kilsby. Rainfall extremes: Toward reconciliation after the battle of distributions. *Water Resources Research*, 50(1):336–352, January 2014. ISSN 00431397. doi: 10.1002/2013WR014211. URL <https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2013WR014211>. Publisher: John Wiley & Sons, Ltd.
- Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. Deep Learning for Precipitation Nowcasting: A Benchmark and A New Model. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/a6db4ed04f1621a119799fd3d7545d3d-Abstract.html>.
- Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, July 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0197-0. URL <https://doi.org/10.1186/s40537-019-0197-0>.

- Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences, October 2019. URL <http://arxiv.org/abs/1704.02685>. arXiv:1704.02685 [cs].
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014. URL <http://arxiv.org/abs/1312.6034>. arXiv:1312.6034 [cs].
- Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise, June 2017. URL <http://arxiv.org/abs/1706.03825>. arXiv:1706.03825 [cs, stat].
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL <http://jmlr.org/papers/v15/srivastava14a.html>.
- Bjorn Stevens and Sandrine Bony. What Are Climate Models Missing? *Science*, 340(6136):1053–1054, May 2013. doi: 10.1126/science.1237554. URL <https://www.science.org/doi/full/10.1126/science.1237554>. Publisher: American Association for the Advancement of Science.
- Ian Strangeways. A history of rain gauges. *Weather*, 65(5):133–138, 2010. ISSN 1477-8696. doi: 10.1002/wea.548. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/wea.548>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/wea.548>.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. Technical report, International Machine Learning Society (IMLS), March 2017. URL <http://arxiv.org/abs/1703.01365>. arXiv: 1703.01365v2 Publication Title: 34th International Conference on Machine Learning, ICML 2017 Volume: 7.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. URL [https://www.cv-foundation.org/openaccess/content\\_cvpr\\_2016/html/Szegedy\\_Rethinking\\_the\\_Inception\\_CVPR\\_2016\\_paper.html](https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html).
- Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. Attention-Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2019. doi: 10.1109/IJCNN.2019.8851881. ISSN: 2161-4407.
- Francisco J. Tapiador, F.J. Turk, Walt Petersen, Arthur Y. Hou, Eduardo García-Ortega, Luiz A.T. Machado, Carlos F. Angelis, Paola Salio, Chris Kidd, George J. Huffman, and Manuel de Castro. Global precipitation measurement: Methods, datasets and applications. *Atmospheric Research*, 104-105:70–97, February 2012. ISSN 01698095. doi: 10.1016/j.atmosres.2011.10.021. URL <https://linkinghub.elsevier.com/retrieve/pii/S0169809511003607>.
- Baijun Tian and Xinyu Dong. The Double-ITCZ Bias in CMIP3, CMIP5, and CMIP6 Models Based on Annual Mean Precipitation. *Geophysical Research Letters*, 47(8):1–11, 2020. ISSN 19448007. doi: 10.1029/2020GL087232.
- Ali Tokay and David A. Short. Evidence from Tropical Raindrop Spectra of the Origin of Rain from Stratiform versus Convective Clouds. *Journal of Applied Meteorology and Climatology*, 35(3):355–371, March 1996. ISSN 1520-0450, 0894-8763. doi: 10.1175/1520-0450(1996)035<0355:EFTRSO>2.0.CO;2. URL [https://journals.ametsoc.org/view/journals/apme/35/3/1520-0450\\_1996\\_035\\_0355\\_eftrso\\_2\\_0\\_co\\_2.xml](https://journals.ametsoc.org/view/journals/apme/35/3/1520-0450_1996_035_0355_eftrso_2_0_co_2.xml). Publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology.



## Bibliography

---

- Benjamin A. Toms, Elizabeth A. Barnes, Imme Ebert-Uphoff, and Imme Ebert-Uphoff. Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. Technical report, Blackwell Publishing Ltd, September 2020. URL <https://onlinelibrary.wiley.com/doi/10.1029/2019MS002002>. arXiv: 1912.01752v2 Publication Title: Journal of Advances in Modeling Earth Systems Volume: 12 Issue: 9 ISSN: 1942-2466.
- Dmitrii Torbunov, Yi Huang, Haiwang Yu, Jin Huang, Shinjae Yoo, Meifeng Lin, Brett Viren, and Yihui Ren. UVCGAN: UNet Vision Transformer cycle-consistent GAN for unpaired image-to-image translation, October 2022. URL <http://arxiv.org/abs/2203.02557>. arXiv:2203.02557 [cs, eess].
- Quang-Khai Tran and Sa-kwang Song. Computer Vision in Precipitation Nowcasting: Applying Image Quality Assessment Metrics for Training Deep Neural Networks. *Atmosphere*, 10(5):244, May 2019. ISSN 2073-4433. doi: 10.3390/atmos10050244. URL <https://www.mdpi.com/2073-4433/10/5/244>. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.
- Dominik Traxl, Niklas Boers, Aljoscha Rheinwalt, and Bodo Bookhagen. The role of cyclonic activity in tropical temperature-rainfall scaling. *Nature Communications*, 12(1): 6732, November 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-27111-z. URL <https://www.nature.com/articles/s41467-021-27111-z>. Number: 1 Publisher: Nature Publishing Group.
- Kevin E. Trenberth, Aiguo Dai, Roy M. Rasmussen, and David B. Parsons. The Changing Character of Precipitation. *Bulletin of the American Meteorological Society*, 84(9):1205–1218, September 2003. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-84-9-1205. URL <https://journals.ametsoc.org/view/journals/bams/84/9/bams-84-9-1205.xml>. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.
- Wen-Ping Tsai, Dapeng Feng, Ming Pan, Hylke Beck, Kathryn Lawson, Yuan Yang, Jiangtao Liu, and Chaopeng Shen. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, 12(1):5988, October 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26107-z. URL <https://www.nature.com/articles/s41467-021-26107-z>. Number: 1 Publisher: Nature Publishing Group.
- Daniele Veneziano, Andreas Langousis, and Pierluigi Furcolo. Multifractality and rainfall extremes: A review. *Water Resources Research*, 42(6), 2006. ISSN 1944-7973. doi: 10.1029/2005WR004716. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2005WR004716>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2005WR004716>.
- Yu Wang, Gu-Yeon Wei, and David Brooks. A Systematic Methodology for Analysis of Deep Learning Hardware and Software Platforms. *Proceedings of Machine Learning and Systems*, 2:30–43, March 2020. URL <https://proceedings.mlsys.org/paper/2020/hash/c20ad4d76fe97759aa27a0c99bff6710-Abstract.html>.
- Zhou Wang and Alan C Bovik. MSE Error : Love It or Leave It ? *IEEE Signal Processing Magazine*, 26(January):98–117, 2009. ISSN 1053-5888. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4775883>. ISBN: 1053-5888.
- Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multi-scale structural similarity for image quality assessment. *Conference Record of the Asilomar Conference on Signals, Systems and Computers*, 2:1398–1402, 2003. ISSN 10586393. doi: 10.1109/acssc.2003.1292216.
- Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. ISSN 1941-0042. doi: 10.1109/TIP.2003.819861. Conference Name: IEEE Transactions on Image Processing.

- Peter A. G. Watson. Machine learning applications for weather and climate need greater focus on extremes. *Environmental Research Letters*, 17(11):111004, November 2022. ISSN 1748-9326. doi: 10.1088/1748-9326/ac9d4e. URL <https://dx.doi.org/10.1088/1748-9326/ac9d4e>. Publisher: IOP Publishing.
- Oliver Watt-Meyer, Noah D. Brenowitz, Spencer K. Clark, Brian Henn, Anna Kwa, Jeremy McGibbon, W. Andre Perkins, and Christopher S. Bretherton. Correcting Weather and Climate Models by Machine Learning Nudged Historical Simulations. *Geophysical Research Letters*, 48(15):e2021GL092555, 2021. ISSN 1944-8007. doi: 10.1029/2021GL092555. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2021GL092555>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021GL092555>.
- Jonathan A. Weyn, Dale R. Durran, Rich Caruana, Yoo Geun Ham, Jeong Hwan Kim, and Jing Jia Luo. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 573(7775):568–572, September 2020. ISSN 1942-2466. doi: 10.1029/2020ms002109. URL <http://www.nature.com/articles/s41586-019-1559-7>.
- Eric M. Wilcox and Leo J. Donner. The Frequency of Extreme Rain Events in Satellite Rain-Rate Estimates and an Atmospheric General Circulation Model. *Journal of Climate*, 20(1):53–69, January 2007. ISSN 0894-8755, 1520-0442. doi: 10.1175/JCLI3987.1. URL <https://journals.ametsoc.org/view/journals/clim/20/1/jcli3987.1.xml>. Publisher: American Meteorological Society Section: Journal of Climate.
- Daniel S. Wilks. *Statistical methods in the atmospheric sciences*. Number v. 91 in International geophysics series. Academic Press, Amsterdam ; Boston, 2nd ed edition, 2006. ISBN 978-0-12-751966-1.
- Daniel S. Wilks and Stéphane Vannitsem. Uncertain Forecasts From Deterministic Dynamics. In *Statistical Postprocessing of Ensemble Forecasts*, pages 1–13. Elsevier, 2018. ISBN 978-0-12-812372-0. doi: 10.1016/B978-0-12-812372-0.00001-7. URL <https://linkinghub.elsevier.com/retrieve/pii/B9780128123720000017>.
- Mark S. Williamson, Chad W. Thackeray, Peter M. Cox, Alex Hall, Chris Huntingford, and Femke J.M.M. Nijsse. Emergent constraints on climate sensitivities. *arXiv*, pages 1–39, 2020. ISSN 23318422. arXiv: 2012.09468.
- Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. DualGAN: Unsupervised Dual Learning for Image-To-Image Translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2849–2857, 2017. URL [https://openaccess.thecvf.com/content\\_iccv\\_2017/html/Yi\\_DualGAN\\_Unsupervised\\_Dual\\_ICCV\\_2017\\_paper.html](https://openaccess.thecvf.com/content_iccv_2017/html/Yi_DualGAN_Unsupervised_Dual_ICCV_2017_paper.html).
- Janni Yuval, Paul A. O’Gorman, and Chris N. Hill. Use of Neural Networks for Stable, Accurate and Physically Consistent Parameterization of Subgrid Atmospheric Processes With Good Performance at Reduced Precision. *Geophysical Research Letters*, 48(6):1–11, 2021. ISSN 19448007. doi: 10.1029/2020GL091363. arXiv: 2010.09947.
- Laure Zanna and Thomas Bolton. Data-Driven Equation Discovery of Ocean Mesoscale Closures. *Geophysical Research Letters*, 47(17):e2020GL088376, 2020. ISSN 1944-8007. doi: 10.1029/2020GL088376. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2020GL088376>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020GL088376>.
- Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. EGSD: Unpaired Image-to-Image Translation via Energy-Guided Stochastic Differential Equations, October 2022. URL <http://arxiv.org/abs/2207.06635>. arXiv:2207.06635 [cs].

## Bibliography

---

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. URL [https://openaccess.thecvf.com/content\\_iccv\\_2017/html/Zhu\\_Unpaired\\_Image-To-Image\\_Translation\\_ICCV\\_2017\\_paper.html](https://openaccess.thecvf.com/content_iccv_2017/html/Zhu_Unpaired_Image-To-Image_Translation_ICCV_2017_paper.html).