On the potential of mapping sea level anomalies from satellite altimetry with Random Forest Regression

Marcello Passaro[1] · Marie-Christin Juhl[1]

Abstract

The sea level observations from satellite altimetry are characterised by a sparse spatial and temporal coverage. For this reason, along-track data are routinely interpolated into daily grids. These grids are strongly smoothed in time and space and are generated using an optimal interpolation routine requiring several pre-processing steps and covariance characterisation. In this study, we assess the potential of Random Forest Regression to estimate daily sea level anomalies. Along-track sea level data from 2004 are used to build a training dataset whose predictors are the neighbouring observations. The validation is based on the comparison against daily averages from tide gauges. The generated dataset is on average 10% more correlated to the tide gauge records than the commonly used product from Copernicus. While the latter is more optimised for the detection of spatial mesoscales, we show how the methodology of this study has the potential to improve the characterisation of sea level variability.

1 Introduction

The monitoring of sea level is conventionally performed using tide gauges and a network of radar altimeters orbiting the Earth. Tide gauges are in situ instruments that register measurements at high frequency (often multiple measurements per hour) and are scattered irregularly along the global coastlines (Woodworth et al. 2016). Altimeters sample along satellite tracks, spanning the same area after a defined number of days depending on the chosen repeating orbit (Fu and Cazenave 2001). Efforts aimed at finding new strategies to improve the characterisation of sea level variability at sub-seasonal time scales and in the coastal and shelf seas to "reduce the gap" between altimetric and tide gauge observations are ongoing as shown in previous works such as Cipollini et al. (2017).

Since altimetry data are along-track measurements scattered in time and space, interpolating algorithms are routinely used to generate sea level maps that are regularly sampled in space and time. The European Union's Earth observation programme, Copernicus, currently releases daily sea level maps and their along-track sources through the Copernicus Marine Service (CMEMS). The CMEMS daily maps are produced using a processing based on optimal interpolation, requiring several steps and assumptions described in Le Traon et al. (1998) and Taburet et al. (2019). The along-track data are sub-sampled and filtered twice, using variable cut-off wavelengths ranging from 200 to 65 km depending on the latitude. The optimal interpolation uses a variable number of observations in time and space, with spatial correlation scales ranging from 80 to 400 km and time correlation scales ranging from 10 to 45 days. It is based on the best linear least square estimator described by Bretherton et al. (1976), in which the covariance matrix of the observations is needed as an input. The covariance matrix is provided by means of assumptions on the errors of the different geophysical corrections applied to the along-track measurements (Pujol et al. 2016).

It has been recently argued that data-driven interpolation is able to perform better than conventional optimal interpolation schemes, whose choice of covariance priors tends

✉ Marcello Passaro
marcello.passaro@tum.de

1 Deutsches Geodätisches Forschungsinstitut der Technischen Universität München, Arcisstraße 21, 80333 Munich, Germany

to over-smooth the sea level variability (Lguensat et al. 2019). The concept behind data-driven interpolation is to exploit machine learning to provide an estimation based on patterns and statistical relations acquired from the training data, rather than from external instructions and assumptions (Zhou et al. 2017). The objective of this paper is to adapt an established machine learning technique to the problem of the estimation of daily sea level maps from along-track altimetry measurements. We use the Random Forest Regression algorithm, described by Breiman (2001), in the implementation of Pedregosa et al. (2011), which has already been successfully used to fill gaps due to missing observations of the ocean (e.g. Gregor et al. (2017) used it to interpolate sparse in situ surface CO_2 observations in the Southern Ocean). In this study, we test a method based on Random Forest Regression to grid sparse along-track measurements based on neighbouring observations. The test is carried out on a regional scale for 1 year of data in the North Sea and we validated using tide gauge data and the optimally interpolated maps from CMEMS. While CMEMS daily grids have only been validated using monthly averages from tide gauges as ground truth, we adopt in this work the daily averages of the Global Extreme Sea Level Analysis (GESLA, version 3), a global archive of high-frequency tide gauge data (Woodworth et al. 2016; Haigh et al. 2021).

2 Data

In this case study, we consider the year 2004 and the extended North Sea including Skagerak/Kattegat in the east and the English Channel in the west. The available altimetry missions in this year were Jason-1, Envisat, Topex/Poseidon and Geosat Follow-on. The North Sea is an ideal testbed for our experiment, thanks to the availability of an extensive

tide gauge network at high frequency, which allows for validation of a daily product. To our knowledge, previous studies involving the comparison of gridded altimetry and tide gauges have only involved monthly data, analysing trends and interannual variability (e.g. Dettmering et al. (2021)). The region of interest and its geographical coordinates are delimited by the red square in Fig. 1.

To train the Random Forest Regression, we use the CMEMS Level 3 (i.e. along-track) sea level anomalies (SLA), reference number: `SEALEVEL_GLO_PHY_L3_REP_OBSERVATIONS_008_062`. We recall that the SLA is defined as the sea level above the mean, corrected for atmospheric and tidal effects. A list of all applied corrections is available in Taburet et al. (2019). We compare the daily machine learning–based SLA from this study (nicknamed ML from now on) with the latest version of the CMEMS Level 4 gridded SLAs, reference number: `SEALEVEL_GLO_PHY_L4_MY_008_047`. We stress the difference in the use of the data sources from CMEMS: the along-track data are the observations that are used to build the regression model; the gridded SLAs are only used for comparison with respect to the results of this study.

As external truth for the validation, we use high-frequency data from tide gauges available from the Global Extreme Sea Level Analysis (GESLA-3, www.gesla.org, Woodworth et al. (2016)). To make the tide gauge data comparable to the altimetry dataset, the following processing steps are needed. Firstly, the atmospheric component is removed using the same correction applied to obtain the SLAs, i.e. the dynamic atmosphere correction from Carrère and Lyard (2003). Secondly, the tidal variability is suppressed using a 40-h LOESS filter, which has been tested to most effectively reduce tidal variance at periods lower than 2 days by Saraceno et al. (2008). The mean of the full sea level record is computed and subtracted from
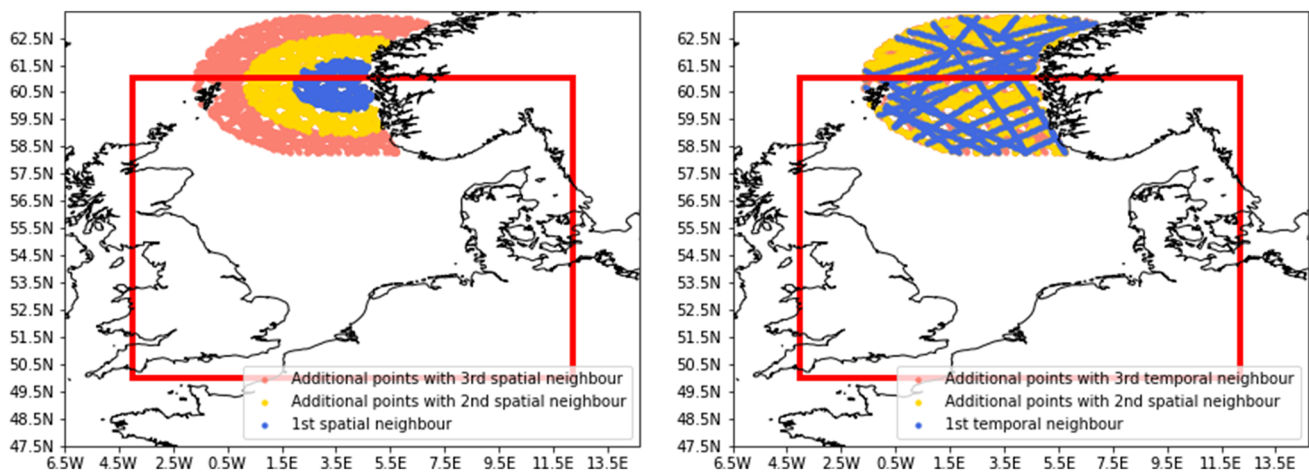

Fig. 1 Examples of along-track observations included in spatial (left) and temporal (right) neighbourhoods associated to one particular location. The red box indicates the area of study. The latter is extended in the search for neighbouring observations, in order to allow for the estimations at the domain's border

each time series in order to obtain the sea level anomalies. Finally, since data are provided at hourly and sub-hourly frequencies, the obtained tide gauge sea level anomalies are averaged at a daily rate.

3 Method

The concept of our methodology is the use of along-track SLAs as truth to train the random forest regressor in the estimation of unknown SLAs (our target variable) on a set of grid points. As predictors, we use means, weighted means and standard deviations of the SLAs at different neighbourhoods in space and time. Furthermore, to better describe the evolution of the target variable in both space and time, the ratios among these predictors from the different neighbours are also used as predictors.

This methodology is inherited from Leirvik and Yuan (2021), who used spatial neighbourhoods to constrain a Random Forest Regression for the interpolation of a surface solar radiation dataset. We expand the methodology by considering the time dimension as well. The following subsections are dedicated to the details of our implementation.

3.1 Preliminary steps

All along-track data for 2004 from CMEMS are collected in the area of study, enlarged by 2.5° in latitude and longitude to guarantee the definition of the neighbourhoods at its borders.

The target variable $y_{training}$ to train the regressor is the field *sla_unfiltered*, where the 1-Hz SLAs (roughly one measurement every 7 km along the track) are stored. The CMEMS Level 4 gridded SLA uses the field *sla_filtered* when interpolating Level 3 data. Such a field is a smoother version of the along-track data obtained using variable filter lengths of several tens of kilometers. Our experiments have shown that the neighbourhood method proposed in this study does not need further filtering and our objective is to keep as much signal as possible. Further discussion and comparison with the CMEMS Level 4 in these regards are provided in Section 4.

We define the locations for computing the SLA, our unknown independent variable y, as the geographical coordinates of a daily unstructured grid. This grid is spaced at intervals of 0.125 degrees in both latitude and longitude, which is equivalent to the grid resolution of the CMEMS Level 4 product.

3.2 Definition of neighbourhoods

We define three spatial neighbourhoods and three temporal neighbourhoods to group the along-track altimetry observations in the proximity of the locations in $y_{training}$ and y.

The spatial neighbourhoods are concentric circles with a radius of 100 km, 200 km and 300 km from the location of the target variable. The temporal neighbourhoods contain the along-track data collected within 5, 10 and 15 days from the time set by the target variable, within a distance that does not exceed 300 km. An example of the along-track locations assembled through the neighbourhoods of one target variable is provided in Fig. 1.

The borders of the neighbourhoods are selected to be within the average global correlation scales of sea level in time and space (see for example Fig. 4 from Pujol et al. (2016)). Nevertheless, the choice for this experimental study is empirical and could be further optimised, for example by using global maps of variable correlation scales depending on the region, such as what is done in the generation of the CMEMS Level 4 grids. We anticipate that we do not observe a substantial change in the performances by slightly changing the neighbourhood definitions.

3.3 Definition of predictors

We define in this section the following classes of predictors: time and space clusters, single-neighbourhoods statistics and multi-neighbourhoods statistics.

3.3.1 Time clustering

The time cluster contains the month in which the variable of interest is defined. Given that the annual cycle is the most prominent SLA periodic signal in time series whose length cannot catch decadal variability, we expect this information to be relevant for the regression. Indeed, Fig. 2a shows the two very different probability densities (PDs) of the SLA for January (blue) and July (red) based on the full training dataset.

3.3.2 Spatial clustering

Several choices could be done concerning spatial clustering. In this exploratory study and in order to generalise the approach, we choose the agglomerative hierarchical clustering (Ward Jr. 1963) in the implementation of Pedregosa et al. (2011). This is an unsupervised classification method that we use to separate the domain in different regions, simply based on the Euclidean distance between the locations in our case. We choose to divide our subdomain in nine clusters, an example of the different PDs of SLA from two of them is visualised in Fig. 2b. We reckon that this is a choice driven by simplicity and other oceanographic information could be used to refine the clustering, for example taking into consideration the spatial correlation with respect to tide gauges (the so-called zone of influence approach from Oelsmann et al. (2020)).

Fig. 2 Probability density of the sea level anomalies associated to specific predictors from the training dataset. Panel (a): months of January and July. Panel (b): two geographical clusters. Panel (c): the mean of the sea level anomalies for the first (100-km radius) and third (300-km radius) spatial neighbourhoods. Panel (d): the mean of the sea level anomalies for the first (5 days) and third (30 days) time neighbourhoods
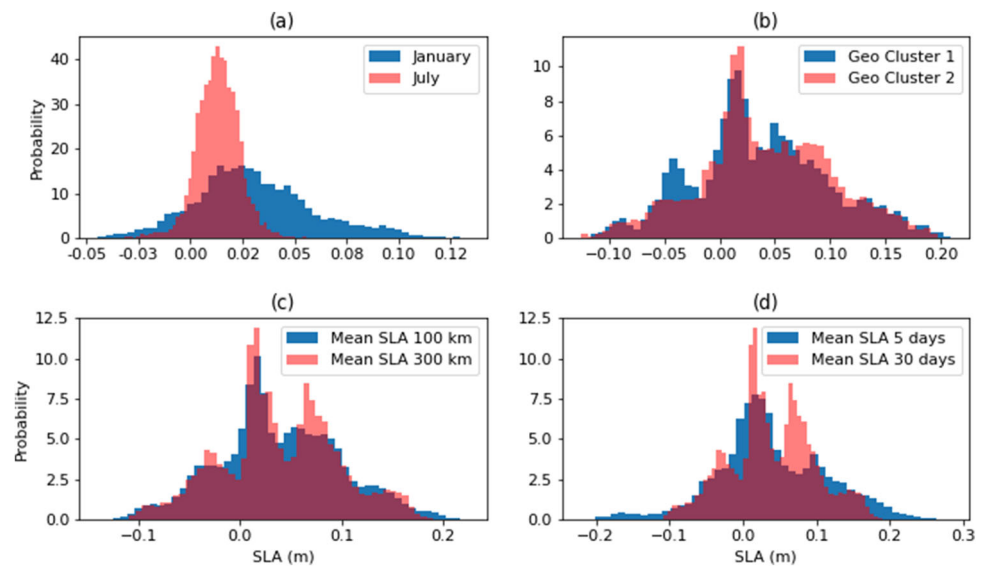
3.3.3 Single-neighbourhoods statistics

For the SLAs contained in every spatial and temporal neighbourhood, we compute the following statistics: mean, spatial-based weighted mean, time-based weighted mean and standard deviation. The weighted means are based on inverse distance weighting, i.e. maintaining the notation of Leirvik and Yuan (2021), the weighted means are defined as:

$$\tilde{z}(N) = \sum_{z_i \in N} \lambda_i z_i \tag{1}$$

where N defines the neighbourhood, z_i is every SLA value within it and the weights λ_i are defined as:

$$\lambda_i = \frac{d_{i0}^{-r}}{\sum_{z_i \in N} d_{i0}^{-r}} \tag{2}$$

For the spatial-based weighted mean, d_{i0} is the Euclidean distance in kilometers of every SLA within the neighbourhood and the location of the target variable. For the time-based weighted mean, d_{i0} is the time difference in seconds between the passing time of the altimeter at the observation i and the time stamp of the target variable. Note that the time difference is multiplied by a factor 10^{-4} in order to achieve similar orders of magnitude between spatial-based and time-based weighted means.

The exponent, r, expresses the relative importance of close-by observations. The highest exponent found in the literature is $r = 5$ from Leirvik and Yuan (2021). We tested lower values and the sensitivity of our results to the choice of r (results are reported in Section 4.2

and found the best performances for $r = 2$). This is not surprising, since a high exponential gives a high importance to the closest observations, while SLA is a field characterised by large spatial and temporal scales of correlation. Given three spatial neighbourhoods and three temporal neighbourhoods, we obtain therefore twenty-four single-neighbourhood predictors. An example of the different PDs of the predictors is given in Fig. 2, where the PD of the mean SLA for the first and third spatial (panel c) and temporal (panel d) neighbourhoods is provided.

3.3.4 Multiple-neighbourhoods statistics

The multiple-neighbourhoods statistics are the ratios between the single-neighbourhoods statistics of the same kind for consecutive neighbourhoods. For example, as in Leirvik and Yuan (2021), considering the mean of the SLAs, we compute the ratio of the mean SLAs between first and second neighbourhoods, $\overline{Z}^{k1,k2}$, and the ratio of the mean SLAs between second and third neighbourhoods, $\overline{Z}^{k2,k3}$. Considering the typical objective of the altimetry missions to achieve a 1-cm SLA accuracy at 1-Hz posting rate (Bonnefond et al. 2013), we round up (or down, for negative numbers) to the nearest centimetre the single neighbourhood statistics previously obtained before computing each ratio.

3.4 Final steps

The predictors are computed for both $y_{training}$ and y locations, generating the predictor matrices $X_{training}$ and X, in which each row corresponds to the predictors associated with each location. Outliers in $y_{training}$ and $X_{training}$ are identified using a 3σ criterion, where σ is the standard

deviation of each variable. Observations in which the SLA or its predictors are identified as outliers are eliminated from the training dataset. Finally, the Random Forest Regression is applied on the training dataset. The obtained regressor $f(\cdot)$ is then applied to estimate the desired SLA on the grid points as $y = f(X)$.

4 Results and discussion

4.1 Examples

To investigate the advantages and the limitations of the generated daily ML product, we first consider examples in time and space. Figure 3 shows the time series of daily averaged data from tide gauges (in green), whose locations are specified at the top of each subplot. The ML product (in blue) and the CMEMS product (in orange), corresponding to the closest location to each tide gauge, are shown for a period comprised between the 15th of January and the 15th of December. This latest choice is because we have only

worked with data from 2004 and therefore the regression would generate worse results at the beginning and the end of the period investigated.

The CMEMS time series appears significantly smooth in time, while the ML product preserves time scales that better match the ones of the tide gauges, although of course the full extent of the high-frequency variability is not captured. Despite CMEMS being smoother than the ML product, the root mean square error (RMSE), computed taking the tide gauges as the truth, is systematically lower for ML. This gives us confidence that the ML time series is not simply noisier than the CMEMS, but it is indeed more accurate.

In Fig. 4, we show a snapshot of ML and CMEMS SLAs for the 24th of April 2004. While the large-scale gradients are similar in both products, the CMEMS map has more defined contours identifying mesoscale variability. The higher variability of ML is the counterpart in space of what has been seen in time in the previous example. The objectives of ML and CMEMS are indeed different: the CMEMS optimal interpolation scheme is dedicated to the retrieval of mesoscale structures (Taburet et al. 2019), while

Fig. 3 Time series estimated from satellite altimetry from this study (ML, blue) and CMEMS (orange) at the closest point to four tide gauges (green), whose coordinates are shown at the top of each panel. Also shown as text is the root mean square error (RMSE) of the altimetry dataset considering the tide gauges as ground truth
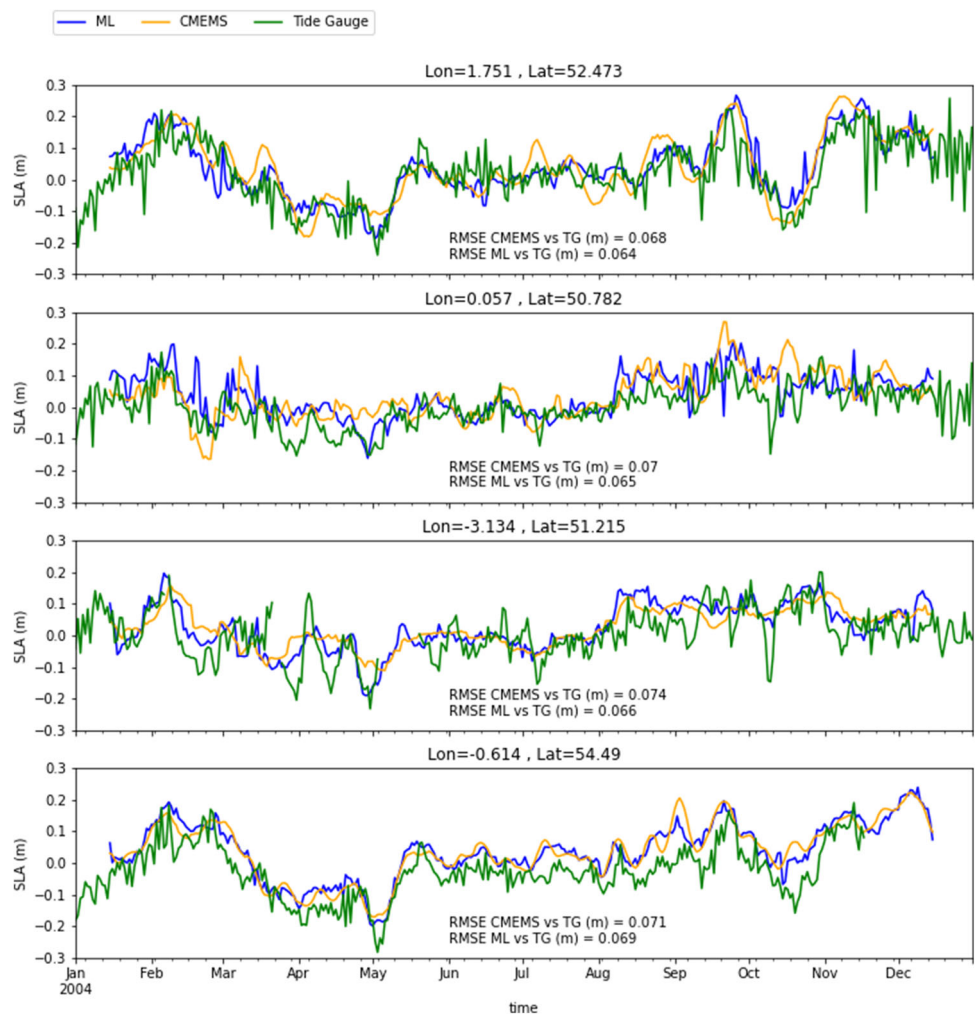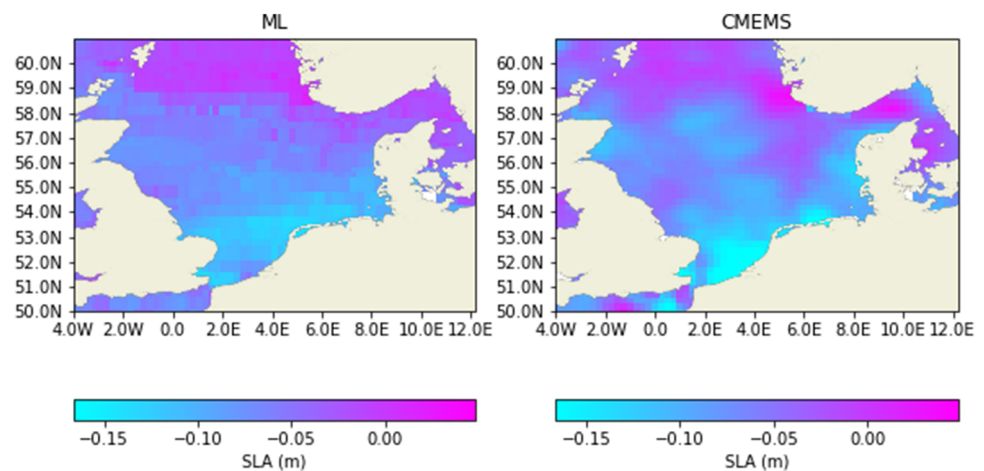
Fig. 4 The daily maps of sea level anomalies (SLA) from this study (ML) and CMEMS estimated for the 24th of April 2004

with ML we attempt to achieve a better compromise to observe local sea level variability. The latest statement is quantified and verified for this case study in Sections 4.2 and 4.3.

The spatial resolution of the ML grid appears degraded compared to CMEMS, in which eddy-like structures can be recognised. Still, it is important to recall that dedicated efforts to assess CMEMS effective spatial resolution concluded that it is not better than 100 km wavelength, which is the resolution reached at the highest latitudes (Ballarotta et al. 2019). Here, we further notice that the CMEMS map is affected by unrealistic extremes of SLA in single pixels in particularly challenging areas such as the English Channel. This is remarkable, considering that the input along-track data of ML and CMEMS are exactly the same, except for the along-track filtering applied by the latter.

4.2 Validation against tide gauges

We assess the general performances of ML and CMEMS computing Pearson's correlation coefficient (CORR) and the RMS between the time series obtained from altimetry and the daily means of the tide gauge data at the closest grid point. Figure 5 shows in the upper panels the RMS and the CORR for ML and in the lower panels the difference with respect to the same statistics computed using CMEMS. The colourbar of the latter is adjusted to show that when the colour is red, an improvement is seen when using the ML product with respect to CMEMS. Good performances (CORR 0.7) are reached along the coasts facing a large open ocean area at the centre of the domain, such as the eastern coast of the United Kingdom (UK). Notably, good performances are also seen in much more enclosed areas situated at the periphery of the domain, such as the Kattegat Sea between Denmark and Sweden (the easternmost part of the domain). This advocates for the robustness of the neighbourhood strategy previously presented. Lowest performances are reached in some enclosed bays and on

both sides of the Channel between UK, France and Belgium (the southernmost part of the domain). Here the quality of the SLAs, also in terms of the geophysical corrections used to extract them, plays a dominant role as shown in previous studies at different temporal scales, such as Dettmering et al. (2021) using monthly time series. The most remarkable result of the validation is that in almost all of the domain (29 tide gauges out of 32) ML performs better than CMEMS. In more than half of the domain, there is at least a 5% improvement in both CORR and RMS considering the tide gauges as ground truth. The average improvement in correlation is 9.98% (6.99% considering RMS), with peaks over 30% that include some of the most problematic areas for satellite altimetry such as the Channel. Concerning the sensitivity of these results to the choice of the weighting factor r, explained in Section 3.3, we tested integer values of r from 1 to 5. The differences between the best and worse average performances are 0.36% in terms of correlation improvement and 1.2% in terms of RMS improvement with respect to CMEMS. In no case among the ones considered did the choice of r determine worse results of ML compared to CMEMS.

In order to understand whether this result depends on the choice of using unfiltered SLAs as training dataset, we also repeated the same experiment starting from the target variable *sla_filtered* (see Section 3.1 for description). We find that using *sla_filtered* produces marginal improvements in the statistics (the average improvement in correlation is 10.17%, 7.45% considering RMS). This corroborates the finding as a result of the ML approach described, with scarce dependence on the smoothing applied to the along-track data. The unfiltered approach is still kept as the basic approach of this study, since our objective is to avoid as much as possible the suppression of the physical signal.

Despite the short time series considered in this experiment, we also compute the magnitude squared coherence (as defined, for example, in Thomson and Emery (2014)) to investigate the agreement between the tide gauges and

Fig. 5 Results of the validation of daily sea level anomaly maps coupled with tide gauges at the closest point. Root mean square (RMS, panel **a**) and Pearson's correlation coefficient (CORR, panel **b**) between the product of this study (ML) and the time series from the tide gauges (panel **a**). Panels **c** and **d**: Difference between these statistics and the equivalent computed using the CMEMS product, in which the red colour palette indicates an improvement using ML
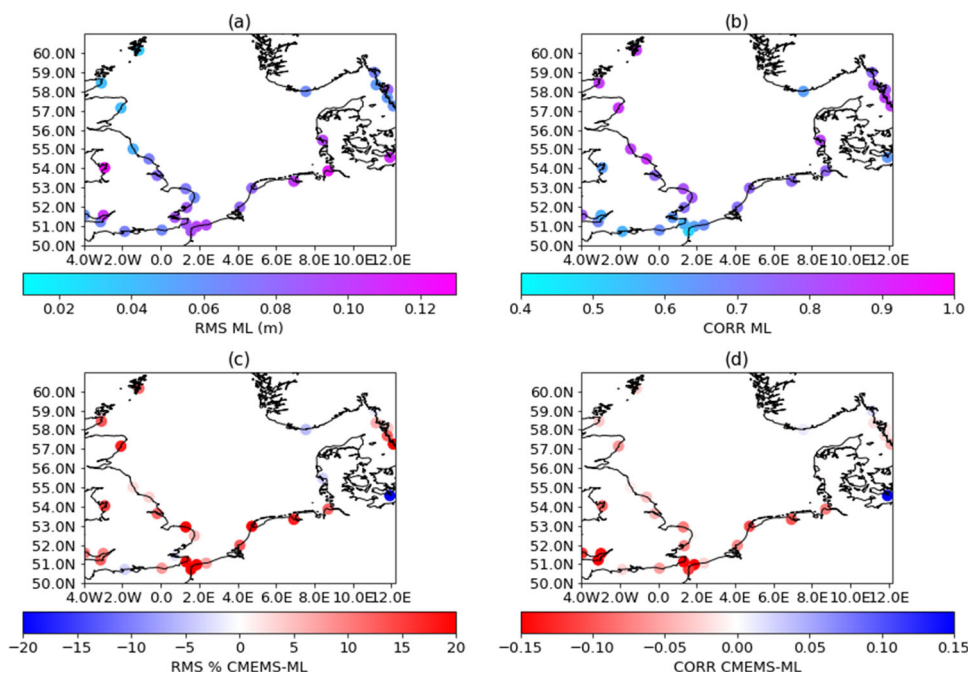
the altimetry time series at different frequencies. We show the mean coherence using ML and CMEMS obtained considering all the available tide gauges and periods below 90 days, in order to have at least 4 time windows to consider out of 1 year of data. The results, displayed in Fig. 6, show that a clear improvement in the coherence is obtained with the ML approach from periods higher than 10 days. Lower periods are dominated by noise, which is expected given that the data are corrected for the dynamic atmosphere correction, which largely suppresses the oceanic variability at these frequencies that are badly sampled by the altimetry constellation (Carrère and Lyard 2003).

4.3 SLA variability

Finally, we assess how realistic the variability of the sea level from the daily grids is. For this purpose, we compute the interquartile range (IQR) of the time series at every grid point and every tide gauge. The IQR is an index of variability computed by taking the difference between the

75th and the 25th percentile of the data, and it is typically used instead of standard deviation or variance because of its robustness. It is commonly used in sea level studies comparing in situ and satellite time series (for example Wöppelmann and Marcos (2016)) and proves fit for our purposes, given that we only assess 1 year of data.

Figure 7a displays the results on the map, showing a consistently increasing variability of ML towards the southeastern part of the domain, which is confirmed by the tide gauge records. In Fig. 7b, the IQR at tide gauges is compared with the variability observed by ML and CMEMS at the closest point. To evaluate this comparison and considering the tide gauges as the ground truth, we compute an index of the average misrepresentation of the sea level variability:

$$Err_{var} = \frac{\sum_{i=1}^{N} \frac{(IQR_{alti} - IQR_{TG})}{IQR_{TG}} \cdot 100}{N} \tag{3}$$

Fig. 6 Results of the validation of daily sea level anomaly maps coupled with tide gauges (TG) at the closest point: mean magnitude squared coherence for periods lower than 90 days expressing similarity between the time series from TG and CMEMS (orange), and from TG and ML (blue)
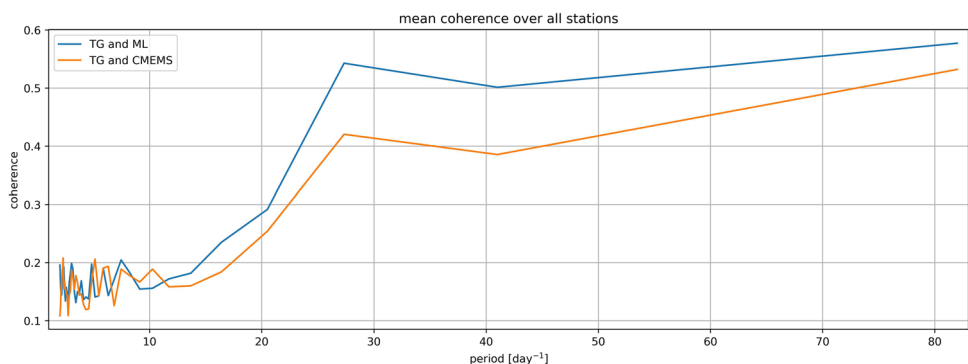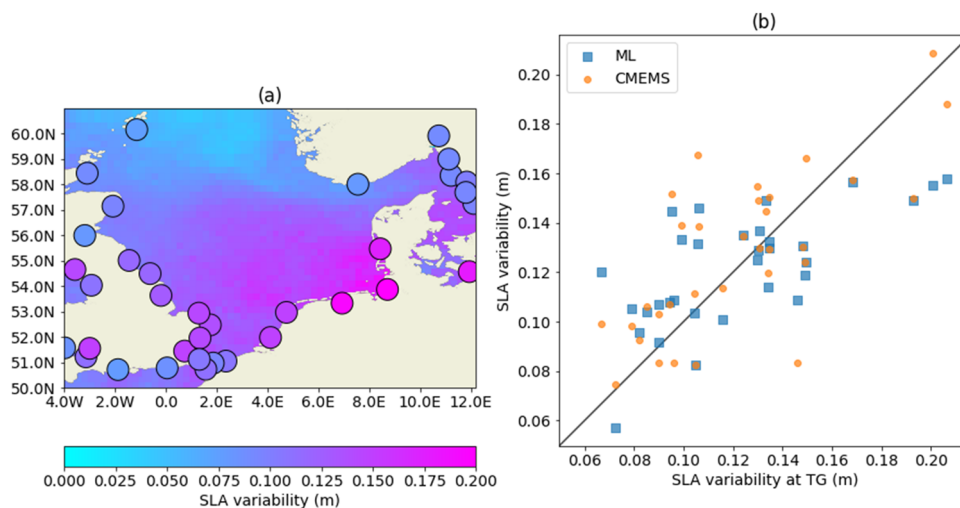
where N is the number of tide gauges and IQR_{alti} is the
IQR of the altimetric time series at the closest point to
each tide gauge. The best results are obtained by ML with
$Err_{var} = 4.4\%$, while when using CMEMS $Err_{var} = 7.6\%$ is achieved. However, it is also noticed that ML
underestimates the variability in the two stations with the
highest IQR.

5 Conclusions

This study has analysed the potential of using a data-
driven approach to produce daily maps of SLAs starting
from along-track observations from satellite altimeters.
This approach allows for circumventing several hypotheses
needed to characterise the covariance of the observations
and their errors in the optimal interpolation. Building on
the existing literature, we have tested a Random Forest
Regression that uses statistics extracted from spatial and
temporal neighbourhoods. By doing so, we have obtained 1
year of daily sea level maps that are on average 10% more
correlated to the observations from tide gauge stations in the
North Sea, compared to CMEMS data.

We believe that the main heritage of this study is the
idea that along-track SLA data can be used to train machine
learning routines aimed at generating gridded maps. The
latter appear less smoothed in space than their CMEMS
counterpart and will therefore need further filtering to be
used for the identification of mesoscale features such as
eddies. Nevertheless, the method presented allows for a
more realistic representation of the sea level variability,
as verified by the comparison against coastal in situ
data. Such comparison has been conducted using high-
frequency tide gauges, which is in our opinion a much

more realistic external validation than the use of monthly
means, if the objective is to assess the capability of the
altimetry constellation to observe sea level at short time
scales.

Since this is an exploratory study, we have to acknowl-
edge both potential and limitations. To speed up the exper-
iments, we have chosen one single year of data (2004), in
which four altimeters were in orbit, and a specific region
(the North Sea). Extending this methodology to a longer
time series will allow to perform coherence studies and
distinguish therefore the performances at different time
frequencies. We have used one single regressor, because
clusters based on time and geographical locations of the
observations were part of the predictors. Nevertheless, the
feasibility of this choice will need to be assessed for stud-
ies involving more years and a wider area, also in terms of
computing time.

The validation against tide gauges shows the strong
potential of machine learning to improve the characterisa-
tion of coastal sea level at a time in which the altimetry
community has recognised the possibilities to improve the
quality of sea level data close to the coast (Benveniste et al.
2020). We expect therefore further improvements by using
SLAs whose estimation is optimised for the coastal zone
(Passaro et al. 2021; Birol et al. 2021), which will neverthe-
less require significant post-processing of the along-track
data, in order not to decrease the quality of the training
dataset.

Declarations

References

Ballarotta M, Ubelmann C, Pujol M-I, Taburet G, Fournier F, Legeais J-F, Faugère Y, Delepoulle A, Chelton D, Dibarboure G et al (2019) On the resolutions of ocean altimetry maps. Ocean Sci 15(4):1091–1109. https://doi.org/10.5194/os-15-1091-2019

Benveniste J, Birol F, Calafat F, Cazenave A, Dieng H, Gouzenes Y, Legeais JF, Leger F, Niño F, Passaro M, Schwatke C, Shaw A (2020) (The Climate Change Initiative Coastal Sea Level Team) Coastal sea level anomalies and associated trends from Jason satellite altimetry over 2002–2018. Sci Data:7. https://doi.org/10.1038/s41597-020-00694-w

Birol F, Léger F, Passaro M, Cazenave A, Niño F, Calafat FM, Shaw A, Legeais J-F, Gouzenes Y, Schwatke C et al (2021) The X-TRACK/ALES multi-mission processing system: new advances in altimetry towards the coast. Adv Space Res 67(8):2398–2415. https://doi.org/10.1016/j.asr.2021.01.049

Bonnefond P, Exertier P, Laurain O, Guinle T, Femenias P (2013) Corsica: a multi-mission absolute calibration site. In: Proceeding of 20 years of progress in radar altimetry, ESA-SP-710

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Bretherton FP, Davis RB, Fandry CB (1976) A technique for objective analysis and design of oceanographic experiments applied to MODE-73. 23(7):559–582. https://doi.org/10.1016/0011-7471(76)90001-2

Carrère L, Lyard F (2003) Modeling the barotropic response of the global ocean to atmospheric wind and pressure forcing-comparisons with observations. Geophys Res Lett 30(6):1275

Cipollini P, Calafat FM, Jevrejeva S, Melet A, Prandi P (2017) Monitoring sea level in the coastal zone with satellite altimetry and tide gauges. Surv Geophys:1–25. https://doi.org/10.1007/s10712-016-9392-0

Dettmering D, Müller FL, Oelsmann J, Passaro M, Schwatke C, Restano M, Benveniste J, Seitz F (2021) North SEAL: a new dataset of sea level changes in the North Sea from satellite altimetry. Earth Syst Sci Data 13(8):3733–3753. https://doi.org/10.5194/essd-13-3733-2021

Fu LL, Cazenave A (eds.) (2001) Satellite altimetry and earth sciences: a handbook of techniques and applications, vol 69, Academic, San Diego

Gregor L, Kok S, Monteiro P (2017) Empirical methods for the estimation of southern ocean CO_2: support vector and random forest regression. Biogeosciences 14(23):5551–5569. https://doi.org/10.5194/bg-14-5551-2017

Haigh ID, Marcos M, Talke SA, Woodworth PL, Hunter JR, Haugh BS, Arns A, Bradshaw E, Thompson P (2021) GESLA: version 3: a major update to the global higher-frequency sea-level dataset. https://doi.org/10.31223/X5MP65

Le Traon PY, Nadal F, Ducet N (1998) An improved mapping method of multisatellite altimeter data. J Atmos Oceanic Tech 15(2):522–534. https://doi.org/10.1175/1520-0426(1998)015<0522:AIMMOM>2.0.CO;2

Leirvik T., Yuan M (2021) A machine learning technique for spatial interpolation of solar radiation observations. Earth Space Sci 8(4):e2020EA001527. https://doi.org/10.1029/2020EA001527

Lguensat R, Viet PH, Sun M, Chen G, Fenglin T, Chapron B, Fablet R (2019) Data-driven interpolation of sea level anomalies using analog data assimilation. Remote Sens 11(7):858. https://doi.org/10.3390/rs11070858

Oelsmann J, Passaro M, Dettmering D, Schwatke C, Sanchez L, Seitz F (2020) The zone of influence: matching sea level variability from coastal altimetry and tide gauges for vertical land motion estimation. Ocean Sci 17:35–37. https://doi.org/10.5194/os-17-35-2021

Passaro M, Müller FL, Oelsmann J, Rautiainen L, Dettmering D, Hart-Davis MG, Abulaitijiang A, Andersen OB, Hoyer JL, Madsen KS, Ringgaard IM, Särkkä J, Scarrott R, Schwatke C, Seitz F, Tuomi L, Restano M, Benveniste J (2021) Absolute Baltic sea level trends in the satellite altimetry era: a revisit. Front Mar Sci. https://doi.org/10.3389/fmars.2021.647607

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Pujol M-I, Faugère Y, Taburet G, Dupuy S, Pelloquin C, Ablain M, Picot N (2016) DUACS DT2014: the new multi-mission altimeter data set reprocessed over 20 years. Ocean Sci 12(5):1067–1090. https://doi.org/10.5194/os-12-1067-2016

Saraceno M, Strub PT, Kosro PM (2008) Estimates of sea surface height and near-surface alongshore coastal currents from combinations of altimeters and tide gauges. J Geophys Res-Space 113(C11):C11013

Taburet G, Sanchez-Roman A, Ballarotta M, Pujol M-I, Legeais J-F, Fournier F, Faugere Y, Dibarboure G (2019) DUACS DT2018: 25 years of reprocessed sea level altimetry products. Ocean Sci 15:1207–1224. https://doi.org/10.5194/os-15-1207-2019

Thomson RE, Emery WJ (2014) Chapter 5 - time series analysis methods. In: Thomson RE, Emery WJ (eds) Data analysis methods in

physical oceanography. 3rd edn., Elsevier, Boston, pp 425–591. third edition edition. ISBN 978-0-12-387782-6. https://doi.org/10.1016/B978-0-12-387782-6.00005-3, https://www.sciencedirect.com/science/article/pii/B9780123877826000053

Ward Jr., JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244

Woodworth PL, Hunter JR, Marcos M, Caldwell P, Menéndez M, Haigh I (2016) Towards a global higher-frequency sea level dataset. Geosci Data J 3(2):50–59. https://doi.org/10.1002/gdj3.42

Wöppelmann G, Marcos M (2016) Vertical land motion as a key to understanding sea level change and variability. Rev Geophys 54(1):64–92. https://doi.org/10.1002/2015RG000502

Zhou Q, Flores A., Glenn NF, Walters R, Han B (2017) A machine learning approach to estimation of downward solar radiation from satellite-derived data products: an application over a semi-arid ecosystem in the US. PloS ONE 12(8):e0180239. https://doi.org/10.1371/journal.pone.0180239