Space Weather®

RESEARCH ARTICLE
10.1029/2023SW003483

Key Points:

- Machine learning-based Vertical Total Electron Content models with 95% confidence intervals (CI) are developed for the first time using four approaches to quantify uncertainties
- Bayesian Neural Network quantifying model and data uncertainties contains ground truth within CIs, but is computationally intensive
- Quantile Gradient Boosting is fastest with comparable performance in terms of uncertainty; CIs largely determined from space weather indices

Supporting Information:

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

Correspondence to:

R. Natras,
randa.natras@tum.de

Citation:

Natras, R., Soja, B., & Schmidt, M. (2023). Uncertainty quantification for machine learning-based ionosphere and space weather forecasting: Ensemble, Bayesian neural network and quantile gradient boosting. *Space Weather*, *21*, e2023SW003483. https://doi.org/10.1029/2023SW003483

Received 9 MAR 2023
Accepted 12 SEP 2023

Author Contributions:

Conceptualization: Randa Natras, Benedikt Soja, Michael Schmidt
Data curation: Randa Natras
Formal analysis: Randa Natras
Funding acquisition: Randa Natras
Investigation: Randa Natras
Methodology: Randa Natras, Benedikt Soja
Software: Randa Natras
Supervision: Michael Schmidt
Validation: Randa Natras
Visualization: Randa Natras
Writing – original draft: Randa Natras

Uncertainty Quantification for Machine Learning-Based Ionosphere and Space Weather Forecasting: Ensemble, Bayesian Neural Network, and Quantile Gradient Boosting

Randa Natras[1], Benedikt Soja[2], and Michael Schmidt[1]

[1]Deutsches Geodätisches Forschungsinstitut (DGFI-TUM), TUM School of Engineering and Design, Technical University of Munich, Munich, Germany, [2]Institute of Geodesy and Photogrammetry, ETH Zurich, Zurich, Switzerland

Abstract Machine learning (ML) has been increasingly applied to space weather and ionosphere problems in recent years, with the goal of improving modeling and forecasting capabilities through a data-driven modeling approach of nonlinear relationships. However, little work has been done to quantify the uncertainty of the results, lacking an indication of how confident and reliable the results of an ML system are. In this paper, we implement and analyze several uncertainty quantification approaches for an ML-based model to forecast Vertical Total Electron Content (VTEC) 1-day ahead and corresponding uncertainties with 95% confidence intervals (CI): (a) Super-Ensemble of ML-based VTEC models (SE), (b) Gradient Tree Boosting with quantile loss function (Quantile Gradient Boosting, QGB), (c) Bayesian neural network (BNN), and (d) BNN including data uncertainty (BNN + D). Techniques that consider only model parameter uncertainties (a and c) predict narrow CI and over-optimistic results, whereas accounting for both model parameter and data uncertainties with the BNN + D approach leads to a wider CI and the most realistic uncertainties quantification of VTEC forecast. However, the BNN + D approach suffers from a high computational burden, while the QGB approach is the most computationally efficient solution with slightly less realistic uncertainties. The QGB CI are determined to a large extent from space weather indices, as revealed by the feature analysis. They exhibit variations related to daytime/nightime, solar irradiance, geomagnetic activity, and post-sunset low-latitude ionosphere enhancement.

Plain Language Summary Space weather describes the varying conditions in the space environment between the Sun and Earth that can affect satellites and technologies on Earth, such as navigation systems, power grids, radio, and satellite communications. The manifestation of space weather in the ionosphere can be characterized using the Vertical Total Electron Content (VTEC) derived from Global Navigation Satellite Systems observations. In this study, the machine learning (ML) approach is applied to approximate the nonlinear relationships of Sun-Earth processes using data on solar activity, solar wind, magnetic field, and VTEC. However, the measurements and the modeling approaches are subject to errors, increasing the uncertainty of the results when forecasting future instances. For reliable forecasting, it is necessary to quantify the uncertainties. Quantifying the uncertainty is also helpful for understanding the ML-based model and the problem of VTEC and space weather forecasting. Therefore, in this study, ML-based models are developed to forecast VTEC within the ionosphere, including the manifestation of space weather, while the degree of reliability is quantified with a target value of 95% confidence.

1. Introduction

Space weather has been identified as a natural hazard to the modern technical infrastructure on which our society is highly dependent. Its accurate and reliable modeling and forecast are therefore essential and rely on modeling nonlinear solar-terrestrial coupling processes. The last few years have witnessed a huge growth in the use of machine learning (ML) and deep learning (DL) to predict complex space weather phenomena, from conditions on the Sun to their effects on Earth (including the ionosphere). Over the next decade, we expect continued rapid development and adaptation of emerging ML/DL tools for operational forecasting systems. However, there is considerable concern about trusting the results of ML/DL models and treating them as a black box because they are difficult to interpret. One of the main issues of previous work is the lack of transparency, as there is no indication when those results should not be trusted, which may lead to scientific skepticism toward ML and DL. Despite their widespread use, there has been little discussion on probabilistic ML/DL and uncertainty quantification (UQ) in the space weather domain. Most studies have focused on providing a single

Writing – review & editing: Randa Natras, Benedikt Soja, Michael Schmidt

prediction/forecast for each input (deterministic), while probabilistic predictions/forecasts have usually not been addressed. The present paper aims to alleviate these issues by extending ML-based models to quantify uncertainty in order to develop the probabilistic model for Vertical Total Electron Content (VTEC) within the ionosphere.

The uncertainty can be classified into two main categories (Abdar et al., 2021; Hüllermeier & Waegeman, 2021; Siddique et al., 2022):

- Model parameter uncertainty: it occurs due to incomplete knowledge, which can be due to a lack of training data or training data information poor. This is the deterministic part of uncertainty, which can be reduced with more knowledge about the system, for example, by adding more information-rich data.
- Data uncertainty: it is related to uncertainty in measurements, which is due to the noise inherent in the data or the stochastic nature of the process generating the data. This is the stochastic part of uncertainty, caused by randomness, and therefore irreducible.

In the ML literature, these uncertainties are often referred to as epistemic and aleatoric uncertainties, respectively (Abdar et al., 2021; Hüllermeier & Waegeman, 2021). The third source of uncertainty relates to the limitation of the learning model to approximate the target function. This is not easy to quantify accurately. For example, model selection involves a particular choice of hyperparameters, and it is impossible to fully explore the hyperparameter space. The choice of hyperparameters can significantly impact the model's accuracy, complexity, and computational cost. Thus, it comes down to a trade-off between the complexity of the model to capture higher-order nonlinear functions and its ability to generalize to unseen data. Ultimately, the ML process consists of various steps of learning and approximating an unknown mapping function from input to output, and the errors and uncertainties associated with these steps may contribute to the uncertainty of the model output.

The most commonly used UQ approaches for ML, in general, are the deep ensembles technique and the Bayesian approximation (Abdar et al., 2021; Kendall & Gal, 2017; Rahaman & Thiery, 2021). There are few examples of UQ studies for estimating a continuous variable in the space weather domain, such as artificial neural network (ANN) with Monte Carlo (MC) dropout as a Bayesian approximation and the negative log-likelihood (NLL) loss function for thermospheric density prediction (Licata & Mehta, 2022), BNN for the geomagnetically induced currents (Siddique et al., 2022), as well as, a least squares-based ensemble of convolutional neural networks (CNN) for the geomagnetic Dst index prediction (Hu et al., 2022). It has been shown that ANN with MC dropout and NLL loss requires much more computational time than ANN with NLL loss and direct probability prediction, but both approaches demonstrated similar accuracy (Licata & Mehta, 2022). Siddique et al. (2022) highlight that estimating uncertainties allows quantifying the degree of reliability of the ML-based model but does not necessarily increase the model accuracy. The least squares-based weighting of the CNN ensemble with a class-balanced cost function was used to account for the imbalance between storm and non-storm cases and provide probabilistic Dst prediction (Hu et al., 2022). The least squares inclusion of both input and output data uncertainties with Bayesian learning using a Long Short-Term Memory neural network resulted in better generalization when applied to the prediction of Earth orientation parameters and Global Navigation Satellite Systems station coordinates (Kiani Shahvandi & Soja, 2022). The initial study of an ML ensemble approach to VTEC forecast in Natras et al. (2022b) showed higher accuracy and improved generalization compared to a single-model approach, with uncertainties estimated as ensemble spread. Other studies on ML-based VTEC modeling and forecasting, such as Y. Han et al. (2022), Kaselimi et al. (2022), L. Liu et al. (2020), Lee et al. (2020), to name a few, have not quantified the uncertainties or provided confidence intervals (CI) of VTEC output, leading to a lack of information on how certain and reliable their ML-based VTEC results are; Natras et al. (2022a) provides an overview of these studies and their results.

Based on the review of existing literature, there has been little discussion on probabilistic ML/DL for VTEC and space weather in general. In this study, we aim to fill this gap by developing and adapting UQ techniques for ML-based VTEC forecasting to produce a probabilistic VTEC model. With this in mind, we analyze and discuss the effectiveness of various techniques for estimating uncertainties and 95% CI of 1-day VTEC forecasting for both quiet and extreme space weather conditions. Section 2 begins with an overview of data preparation, then describes four methods for estimating uncertainty, and ends with an outline of models and hyperparameters optimization. Section 3 provides a detailed analysis of the UQ models for test case studies. Our conclusions are drawn in the final section.

2. Methodology

2.1. Data

This study deals with supervised learning, in which a set of both input and output data is clearly specified and prepared, called training data, which is needed to learn the function that maps the input variables to an output variable. A training sample consists of the vector $\mathbf{x}_i$ and an output $y_i = F(\mathbf{x}_i)$ with $i = \{1, 2, \ldots, N\}$. The vectors $\mathbf{x}_i$ can be interpreted as the rows of the $N \times P$ predictor matrix $\mathbf{X} = \left(\mathbf{X}_i^T\right)$, whereas the columns represent the input features $\widetilde{\mathbf{x}}_p$ with $p = \{0, 1, 2, \ldots, P-1\}$ (Natras et al., 2022a).

In this study, VTEC is obtained from Center for Orbit Determination in Europe (CODE) global ionospheric maps (GIM), computed via spherical harmonics up to degree 15 (Schaer, 1999), and interpreted as GT. Because the temporal resolution of the CODE GIM was updated from 2 hr to 1 hr in 2015, we used data starting from January 2015 to develop the VTEC model with 1-hr intervals. September 2017 was an extremely active space weather period, with the Sun emitting 27 M-class and 4 X-class flares, as well as several earthward-directed coronal mass ejections (CME) (https://www.nasa.gov/feature/goddard/2017/september-2017s-intense-solar-activity-viewed-from-space). Therefore, the year 2017 is selected for testing, and data from January 2015 to December 2016 are used for training and cross-validation, The training set consists of a total of 17,544 samples, while the test set contains 8,760 samples. To model the solar-terrestrial processes and the impact of space weather on VTEC, data on solar and geomagnetic activity were downloaded from the OMNIWeb NASA Service and added as input features. The data set is prepared with a 1-hr resolution, denoted D1, corresponding to Table 1 of Natras et al. (2022a). It consists of the following input features of $\mathbf{x}_i$ at timestamp i:

(a) VTEC for grid points at 10° of longitude, and 10°, 40°, and 70° of latitude;
(b) OMNIWeb data: sunspot number, F10.7 solar radio flux, solar wind plasma speed, interplanetary magnetic field Bz index, geomagnetic field (GMF) Dst index, GMF Kp index, auroral electrojet (AE) index;
(c) Derived VTEC features: exponential moving average (EMA) of VTEC over the previous 30 and 4 days, first and second VTEC derivatives;
(d) Hour of the day (HoD) and day of the year (DoY),

and the output $\mathbf{y}_i = \mathbf{VTEC}(i + 24)$ for the 1-day forecast. Grid points for VTEC were selected along the same longitude (10°) to represent VTEC latitudinal variations alongside other VTEC variability. Separate models were developed for each grid point.

The input data for artificial neural networks were standardized to obtain data with a mean of zero and a standard deviation of one. Learning algorithms based on decision trees (Sections 2.2.1 and 2.2.2) do not require data normalization since they are not sensitive to the scale of input features, and the data were not standardized in these cases. Moreover, a neural network benefits from transforming time information to preserve its cyclic significance:

$$HoD_{\sin} = \sin\left(\frac{2\pi \cdot HoD}{24}\right), \quad HoD_{\cos} = \cos\left(\frac{2\pi \cdot HoD}{24}\right)$$

$$DoY_{\sin} = \sin\left(\frac{2\pi \cdot DoY}{365.25}\right), \quad DoY_{\cos} = \cos\left(\frac{2\pi \cdot DoY}{365.25}\right). \tag{1}$$

2.2. Methods

In the following, we present different approaches to determine model and data uncertainties in ML-based UQ VTEC models.

2.2.1. Ensemble Approach

Ensemble modeling combines multiple diverse models to predict an outcome using either different algorithms or different data sets. The ensemble model, called Super-Ensemble (SE) (Natras et al., 2022b), aggregates the mean result across all base models to produce a final prediction with reduced generalization error. This approach improves the prediction compared to the single base model within the ensemble by averaging the results over a set of functions of well-performing models (Natras et al., 2022b). In this study, ensemble modeling combines three learning algorithms, namely Random Forest (Breiman, 2001), Adaptive Boosting (AdaBoost) (Freund &

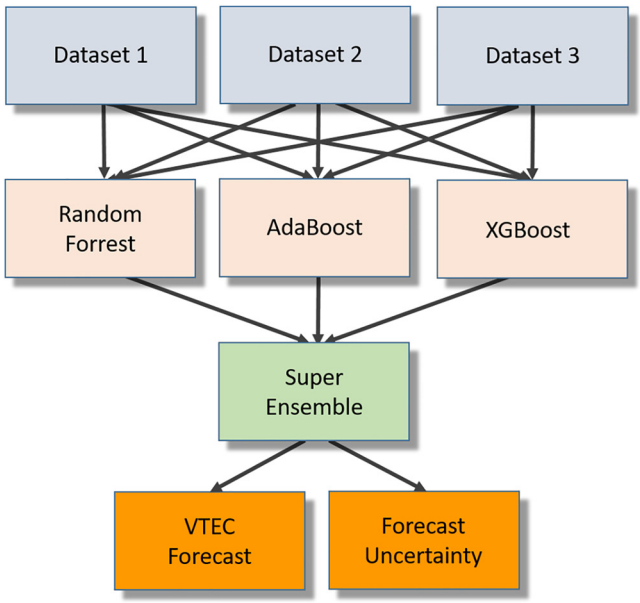

Figure 1. Flowchart of the ensemble modeling procedure trained on three different data sets using the three different learning algorithms Random Forest, AdaBoost, and eXtreme Gradient Boosting (XGBoost). After the nine individual model runs, the results are combined into a Super-Ensemble model, which provides the Vertical Total Electron Content forecast values and their uncertainty.

Schapire, 1997) and Gradient Boosting (Friedman, 2001) on three data sets consisting of different versions of input features and output. These algorithms are based on decision tree learning but follow different computation strategies. Random Forest belongs to the bagging approach of learning many diverse random trees, while the other two algorithms are boosting approaches of sequential learning that aim to reduce the errors of the tree from the previous step. Moreover, boosting is realized differently in AdaBoost by assigning different weights to observations depending on the model performance in the previous step and in Gradient Boosting by training the models on the gradient of the objective cost function of the previous step. For more details, see Natras et al. (2022a).

In addition to training the ensemble members with different learning algorithms, further randomness is introduced into the ensemble by training with different versions of the data set to increase the number of ensemble members and increase the diversity between them, as shown in Figure 1. Therefore, we created three sets of data from the D1 data set introduced in Section 2.1:

1. Data set D1 with $\mathbf{x}_i$, y_i for $i = 1, 2, \ldots, N$;
2. Daily differences for the input features and output: The data, except HoD and DoY, are time-differenced with $\Delta\mathbf{x}_i$, Δy_i by calculating the difference between an observation at time step $i + 24$ and observation at time step i so that $\Delta\mathbf{x}_i = \mathbf{x}_{i+24} - \mathbf{x}_i$ and $\Delta y_i = y_{i+24} - y_i$. The EMA and time derivatives of VTEC are calculated from the differenced VTEC values. At the end, the VTEC forecast is reconstructed from the forecasted VTEC daily difference by adding the VTEC value from 24 hr ago.
3. The input of the data set from point 1 and the input of the daily differenced data set from point 2 are used as input features, while the output comes from the data set from point 1.

Daily differences remove the dominant daily VTEC variations so that the model can learn the remaining signatures associated with other sources of VTEC fluctuations. Such a data strategy demonstrated improved generalization and accuracy of 1-day VTEC forecasting in ensemble tree learning (Natras et al., 2022a), as well as in the convLSTM VTEC model (L. Liu et al., 2022). In addition, differencing reduces temporal dependencies and trends and stabilizes the mean of the data set, which can improve modeling.

The employed cost function is the mean squared error (MSE), defined as

$$Cost_m = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}\left(y_i, \hat{F}(\mathbf{x}_i)\right) = \frac{1}{N}\sum_{i=1}^{N} e_i^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_{m_i}\right)^2, \tag{2}$$

where $\mathcal{L}$ is the loss function, y_i is the GT VTEC, $\hat{F}(\mathbf{x}_i)$ is an approximation function of the function $F(\mathbf{x}_i)$ that maps the input $\mathbf{x}_i$ to the output y_i, and $\hat{y}_{m_i}$ is the VTEC forecast of the mth model with $m = \{1, 2, \ldots, M\}$.

The ensemble approach can be viewed as an approximation of a distribution, and thus, its diversity can be used as an indicator of the model parameter uncertainty (Hüllermeier & Waegeman, 2021). In this case, the results of M independently trained models are averaged, forming a joined distribution $p(\mathbf{y}|\mathbf{X})$ as

$$p(\mathbf{y}|\mathbf{X}) = \frac{1}{M}\sum_{m=1}^{M} p(\mathbf{y}|\mathbf{X}, \theta), \tag{3}$$

where θ represents a set of model parameters. Nine models, $M = 9$, are developed for each of the 3 VTEC grid points, resulting in a total of 27 models. The randomness in the nine models in this study is introduced by the learning algorithms and the data. More specifically, by training the three algorithms mentioned above on each of the three data sets individually. The final output $\hat{y}_i$ is estimated as the ensemble mean μ_i

$$\hat{y}_i = \mu_i = \frac{1}{M}\sum_{m=1}^{M}\hat{y}_{m_i} \tag{4}$$

and the standard deviation across the ensemble for observation time i is defined as

$$\sigma_i = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{y}_{m_i} - \hat{y}_i\right)^2}. \tag{5}$$

The standard deviation of the ensemble members with respect to the ensemble mean, known as the ensemble spread, provides an estimate of the uncertainties. The ensemble spread is represented as a probabilistic prediction in terms of lower bounds (LB) and upper bounds (UB) with 95% confidence, defined by

$$UB = \hat{y}_i + 2\sigma_i, \quad LB = \hat{y}_i - 2\sigma_i. \tag{6}$$

2.2.2. Quantile Gradient Boosting

Quantile methods (Koenker & Hallock, 2001) can be seen as an extension of classical least squares model estimation for the conditional mean function to the estimation of models for the conditional median function and the full range of other conditional quantile functions. The quantile function does not require a specification of variance changes and can thus model heterogeneous variation in the objective loss distribution (Chan, 2021). Moreover, this approach avoids the distributional assumption, that is, it does not assume a Gaussian error distribution (unlike most traditional methods) and can be used when the error distribution is non-Gaussian (Chan, 2021). Quantiles can be estimated by multiplying different quantile values β by positive and negative residuals in the loss function to obtain the quantile loss as

$$\begin{aligned} \mathcal{L}(e_i|\beta) &= \begin{cases} \beta \cdot e_i & \text{if} \quad e_i \geq 0, \\ (\beta - 1) \cdot e_i & \text{if} \quad e_i < 0 \end{cases} \qquad e_i = y_i - \hat{y}_i \\ Cost(\mathbf{e}|\beta) &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(e_i|\beta). \end{aligned} \tag{7}$$

The quantile values of β are set to 0.025 and 0.975 for estimating the lower and upper confidence bounds, respectively, to obtain a CI of 95%. The mean quantile $\beta = 0.50$ provides the median VTEC forecast. To estimate other CI levels of 90% and 99%, the quantile values must be changed to $\beta = \{0.05, 0.95\}$ and $\beta = \{0.005, 0.995\}$, respectively.

Quantile loss has been shown to model data uncertainty in neural networks (Amell et al., 2022; Tagasovska & Lopez-Paz, 2019). In this study, we applied quantile loss with a Gradient Boosting tree. The Gradient Boosting algorithm and its implementation for VTEC forecast are explained in Natras et al. (2022a). We chose Gradient Boosting because it is fast (Natras et al., 2022a), performs well on structured input data even for relatively small data sets (Duan et al., 2020), and has proven to be a powerful method in many data science competitions (Chen & Guestrin, 2016). Moreover, Vasseur and Aznarte (2021) compared the performance of 10 ML algorithms with quantile loss for predicting NO_2 pollution and found that Gradient Boosting outperformed the other models with better results for all metrics examined.

2.2.3. Bayesian Neural Network

The Bayesian neural network (BNN) represents a modification of an ANN in which the deterministic network parameters or weights are replaced by probability distributions of those weights (Abdar et al., 2021; Blundell et al., 2015; Kendall & Gal, 2017); for more details on the architecture and computation of an ANN, see Natras, et al. (2023a). The probability distributions are used to model the uncertainty in the weights and consequently can be used to estimate the uncertainty due to the model parameter uncertainty based on Bayes' theorem. The posterior parameters θ to be trained are the mean μ and standard deviation σ of the posterior weight distribution. They can be learned by variational Bayesian inference during the training process, facilitated by a standard neural network backpropagation technique during the training process (Blundell et al., 2015). That technique is called Bayes by Backprop and is implemented in this study.

Given a training data set $D = (x_i, y_i)$ with $i = 1, 2, \ldots N$, the likelihood function $p(D|w)$ can be constructed, which is a function of the weights w. Maximizing the likelihood function yields the maximum likelihood estimate of w. The usual optimization objective in ML training is to minimize the NLL. Multiplying the likelihood by a

prior distribution $p(w)$ is proportional to the posterior distribution $p(w|D) \propto p(D|w)p(w)$ according to Bayes' theorem (Koch, 2018). An analytical solution for the posterior $p(w|D)$ in neural networks is not feasible. We can approximate the true posterior with a variational distribution $q(w|\theta)$ of the function whose parameters we want to estimate. This can be done by minimizing the Kullback-Leibler (KL) divergence between $q(w|\theta)$ and the true posterior $p(w|D)$.

KL divergence measures how close the variational probability distribution of the weights $q(w|\theta)$ is to the posterior probability distribution of the weights $p(w|D)$. It is also called relative entropy in probability and information theory (Murphy, 2012). Normally, the reverse KL divergence is used

$$\begin{aligned} KL(q(w|\theta)\|p(w|D)) &= q(w|\theta)\cdot \log\frac{q(w|\theta)}{p(w|D)} \\ &= -\log p(D|w) + KL[q(w|\theta)\|p(w)]. \end{aligned} \tag{8}$$

The idea behind variational inference is to choose an approximation $q(w|\theta)$ to the distribution and then try to make this approximation as close as possible to the true posterior $p(w|D)$. This reduces variational inference to an optimization problem, and from Equation 8, the objective cost function can be defined as follows

$$Cost = \frac{1}{N}\sum_{i=1}^{N}(-\log p(D|w) + KL[q(w|\theta)\|p(w)]), \tag{9}$$

which can be split into two parts: the left term of the loss function on the right side corresponds to the NLL, and the right term is the KL divergence between the variational distribution $q(w|\theta)$ and the prior $p(w)$, which can also be seen as the regularization term.

The prior weight distribution is a Gaussian distribution with a mean $\mu = 0$ and a diagonal covariance with a standard deviation $\sigma = 1$. A sample of the weights w is obtained by randomly sampling ϵ from $\mathcal{N}(0, 1)$, then scaling it by a standard deviation σ, and shifting it by a mean μ as

$$w = \mu + \sigma\cdot\epsilon. \tag{10}$$

For numerical stability, the network is parametrized with ρ instead of σ. ρ is transformed with the so-called softplus activation function as

$$\sigma = \log(1 + \exp(\rho)) \tag{11}$$

to ensure that σ is always non-negative (Blundell et al., 2015). The algorithm proceeds by sampling from the variational posterior distribution, computing a forward pass through a network, and then backpropagating through the model parameters to update them. The gradients are calculated with respect to the mean and the standard deviation to update the previous distribution parameters using the stochastic gradient descent optimization algorithm (Bottou, 2012). The parameters are updated stepwise, controlled by the learning rate, along a preferred direction, which is a function of the previous gradient.

The Gaussian likelihood is assumed in this study, parameterized by the mean and standard deviation as

$$p(D|w) = l(y|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}. \tag{12}$$

The NLL loss is defined as

$$\begin{aligned} \mathcal{L} &= -\log l(y_i|\mu,\sigma) \\ &= \frac{1}{2}\left[\log\left(\sigma^2\right) + \frac{(y_i-\mu)^2}{\sigma^2} + \log(2\pi)\right] \\ &= \frac{1}{2}\left[\log\left(\sigma^2\right) + \frac{(y_i-\mu)^2}{\sigma^2} + C\right], \end{aligned} \tag{13}$$

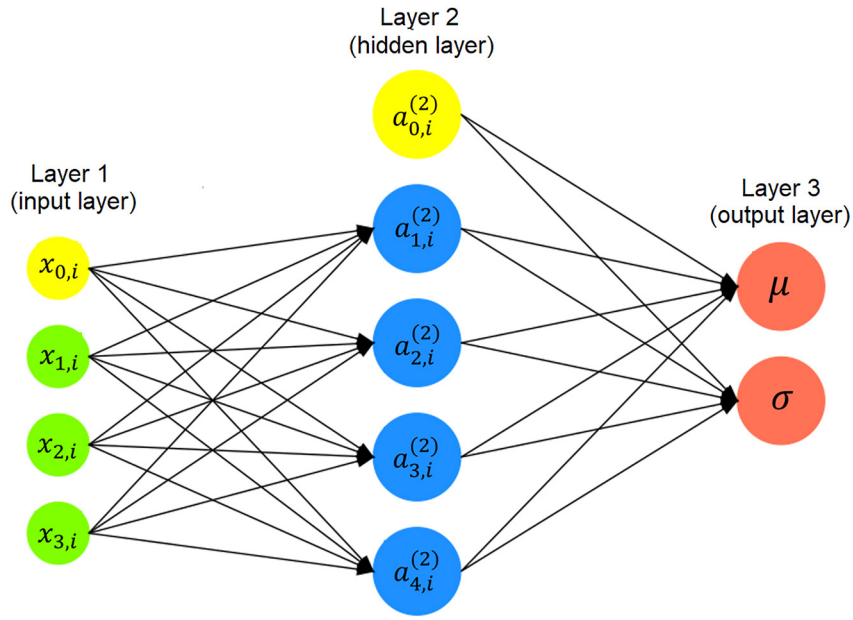

Figure 2. Simplified representation of the Bayesian neural network (BNN) architecture with probabilistic implementation to output the probability distribution parameters: μ and σ. The yellow circles represent the bias neurons. The green circles are the input neurons (simplified here to 3, but 14 in the BNN + D VTEC model), the blue circles are hidden neurons (4 here, but 32 in the BNN + D VTEC model), and the orange circles are the output neurons.

where y_i is the observed value or GT, μ is the predicted mean, and σ is the standard deviation. C is a constant equal to $\log(2\pi)$, which can be neglected. The loss function in Equation 13 is also known as the negative logarithm of predictive density (Licata & Mehta, 2022).

For BNN, we assume a fixed data noise, as usual. Then the loss in Equation 13 corresponds to the squared error loss, and the left term on the right side in Equation 9 becomes the standard MSE cost, similar to Equation 2. This Bayesian approach to an ANN aims to capture the model parameter uncertainty due to limited training data. Each time the BNN model is run with the same input variables, a new set of parameters is sampled from the distribution, and a result is produced. In this study, the VTEC forecast is estimated as the mean of an ensemble of results from 100 iterations, while the 95% CI is calculated as in Equation 6.

The BNN implementation described so far is deterministic, that is, it produces a single VTEC forecast for each run, and the uncertainty is calculated from an ensemble of many iterations. The BNN can be extended to a probabilistic implementation by enabling the model to output a distribution and quantify the data uncertainty. In this case, the data noise is assumed to be data-dependent rather than fixed, and thus, it is learned as a function of the data. Therefore, the NLL from Equation 13, which accounts for the observation noise, is used in Equation 9 to compute how likely the GT values are to deviate from the estimated distribution produced by the model. The model can then provide a probability distribution as an output, that is, μ and σ, instead of a single point estimate. To provide μ and σ as output values, a custom output layer is created with two neurons, shown in Figure 2: one for mean output and one for standard deviation output. The 95% CI is computed from the predicted standard deviation according to Equation 6.

All approaches used in this study are summarized in Table 1. The 95% CI in the SE, BNN, and BNN + D approaches is approximated by multiplying the standard deviation by 2. If necessary, other CI can also be estimated. For example, multiplying the standard deviation by 1.64 gives a 90% confidence level, and by 2.58 gives a 99% confidence level. For the QGB approach, the quantile values must be adjusted accordingly to estimate 90% and 99% CI, as already mentioned in Section 2.2.2.

2.3. Models Optimization and Hyperparameters

Optimization of a ML model includes adjusting the hyperparameters to minimize the objective cost function. In this study, the hyperparameters were tuned using 20-fold time-series cross-validation (Natras et al., 2022a) and

Table 1
Approaches of Applying Different Uncertainty Quantification Methods on Different Learning Algorithms and Their Abbreviations

Approaches	Cost/Loss	Learning algorithms	Abbreviations
Ensemble modeling: Super-Ensemble	MSE	Bagging and Boosting	SE
Quantile Gradient Boosting	Quantile	Gradient Boosting	QGB
Bayesian inference	MSE + KL	Bayesian Neural Network	BNN
Bayesian inference	NLL + KL	Bayesian Neural Network	BNN + D

grid search within the hyperparameter range, see Table S1 in Supporting Information S1. Table 2 summarizes the selected values for the hyperparameters.

3. Results

The analysis is performed for the year 2017 (1 January–31 December 2017), for a period with space weather events (6–10 September 2017), and for a quiet period concerning solar and geomagnetic activity (25–29 April 2017). Figures 3 and 4 show the 1-day VTEC forecast in orange with a 95% CI in green using the SE and BNN approaches, and the QGB and BNN + D approaches, respectively, for the quiet period (left) and the storm period (right) in 2017. The results of mean/median VTEC from different ML-based UQ VTEC models are summarized in Table 3.

As the baseline models, we use the frozen ionosphere and the Multi-Layer Perceptron (MLP) model. For the frozen ionosphere, we define that VTEC(i + 24) equals VTEC(i), that is, we assume the state of the frozen ionosphere with respect to the previous day. This assumption is consistent with the prevailing diurnal VTEC variability, where the next day's VTEC should not be significantly different from the previous day's VTEC under quiet conditions. The MLP model is the classical type of neural network and represents a fully connected ANN consisting of one or more hidden layers of neurons. MLP is the most commonly used ML method for VTEC modeling and forecasting (e.g., Ferreira et al., 2017; Orus Perez, 2019; Özkan, 2022). The International Reference Ionosphere (IRI) 2016 is used as a third baseline, where VTEC was extracted at the height of 450 km, and the upper height for TEC integration was set at 20,000 km. The IRI analysis was conducted for two study periods: one in April and a second in September 2017, and the detailed analysis is shown in Figure 5.

The period 6–10 September 2017, represents the most intense solar activity period with the strongest solar flare of X9.3 class, which peaked at 12:02 UT on 6 September. Earthward-directed CMEs were emitted from the Sun on 4 and 6 September (Imtiaz et al., 2020). The first CME arrived at about 23:43 UT on 6 September and caused moderate geomagnetic conditions on 7 September, while the second CME from the X9.3 solar flare triggered a sudden storm commencement at 23 UT on 7 September. This resulted in severe geomagnetic storms on 8 September with a maximum Kp = 8 (Figure 4, bottom). The main phase of the storm was characterized by the two pronounced minima of the Dst index at around 1 and 14 UT on 8 September. Thereafter, the recovery phase began and lasted for about 3 days, that is, until 11 September.

The SE and BNN methods provide narrow CI ranging from less than 1 TECU to 2 TECU from the mean VTEC forecast, as shown in Figure 3. However, about 50% of GT VTEC values in 2017 are outside their 95% CI in Table 3. This indicates that the approaches that capture only the model uncertainties produce over-confident VTEC CI. During the disturbed space weather period in September 2017, the GT is outside the CI by up to 4

Table 2
Selected Hyperparameters (hl Stands for Hidden Layer and lr for Learning Rate)

Model	Hyper-parameters
SE	see Table 2 in Natras et al. (2022a)
QGB	tree depth = {3, 4, 5}, number of trees = {50, 100, 150}, lr = 0.1
BNN	batch = 500, epoch = {500, 1,000}, lr = 0.001, 1 hl with 32 neurons
BNN + D	batch = 500, epoch = 2,000, lr = {0.01, 0.1}, 1 hl with 32 neurons

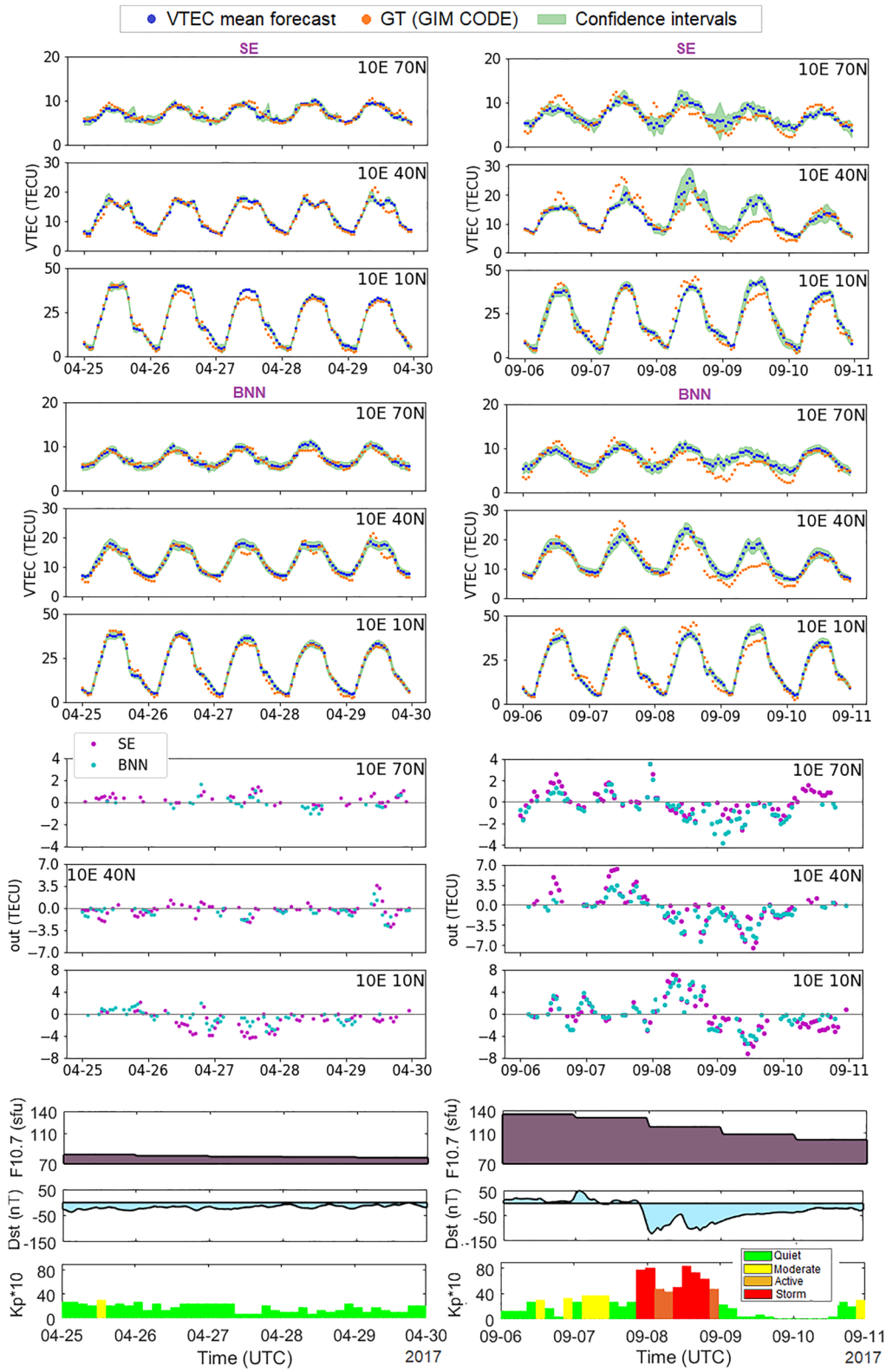

Figure 3. Mean Vertical Total Electron Content (VTEC) forecast and 95% confidence intervals (CI): SE (first panel) and Bayesian neural network (second panel) for selected grid points. Third panel: ground truth (GT) VTEC outside CI (positive value: the amount by which GT is higher than upper CI, negative value: the amount by which GT is lower than lower CI). Fourth panel: indices of F10.7, Dst, and Kp · 10 ($Kp < 3$, $3 \leq Kp < 4$, $4 \leq Kp < 5$, and $Kp \geq 5$ denote quiet, moderate, active and storm conditions, respectively). Left: 25–29 April 2017, right: 6–10 September 2017.

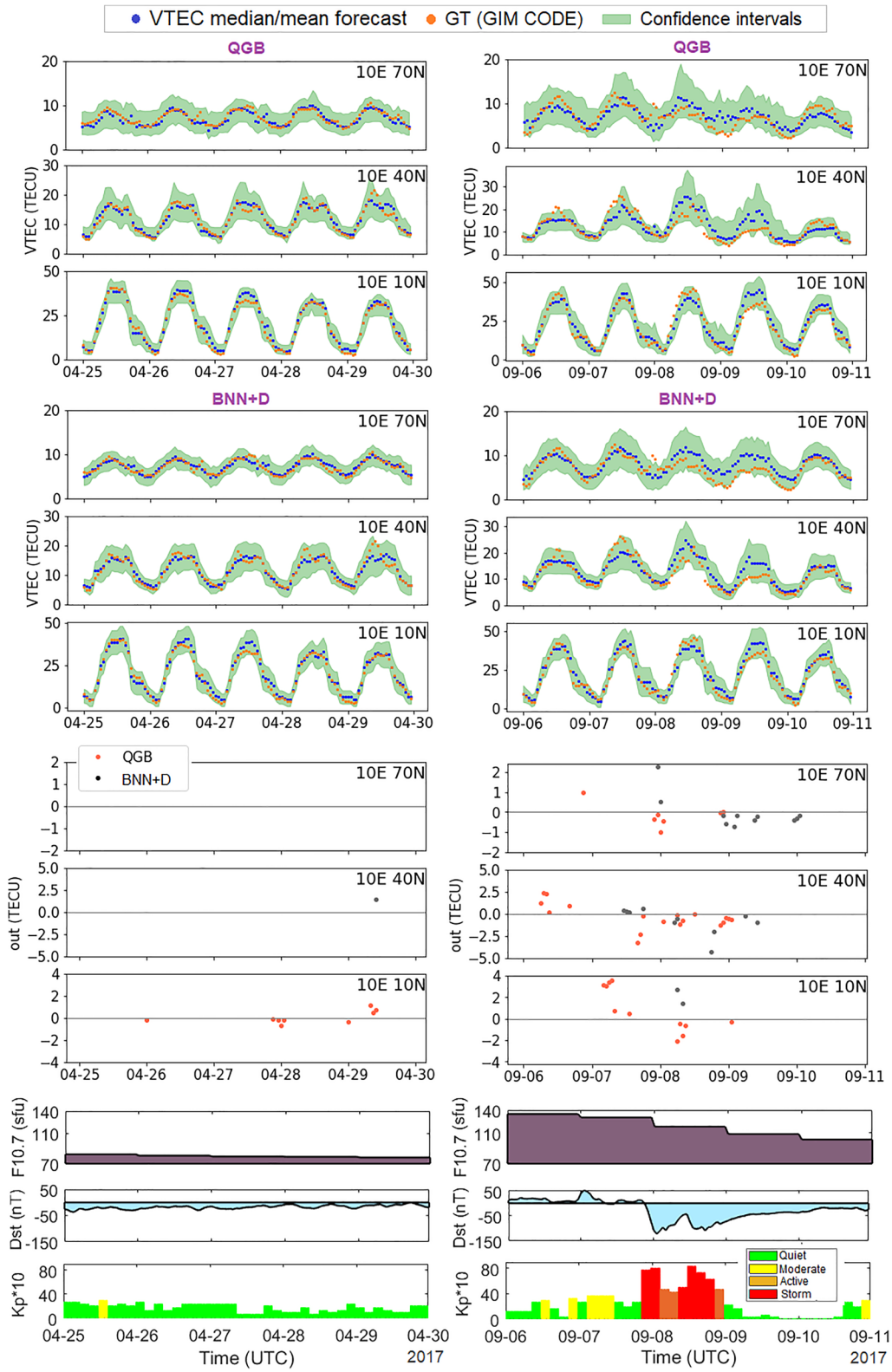

Figure 4. Median/mean Vertical Total Electron Content (VTEC) forecast and 95% confidence intervals (CI): Quantile Gradient Boosting (first panel) and BNN + D (second panel) for three selected grid points. Third panel: ground truth (GT) VTEC outside CI (positive value: amount by which GT is higher than upper CI limit, negative value: amount by which GT is lower than lower CI limit). Fourth panel: indices of F10.7, Dst, and Kp · 10. Left: 25–29 April 2017; right: 6–10 September 2017.

Table 3
Statistics on the Test Data Set for 1-Day Probabilistic Vertical Total Electron Content Forecast

	1 January–31 December 2017	6–10 September 2017	25–29 April 2017
Model	RMS, Corr., CI_{avg}, In(%)	RMS, Corr., CI_{avg}, In(%)	RMS, Corr., CI_{avg}, In(%)
VTEC:10°70°			
SE	1.03, 0.92, 0.74, 51.89	1.73, 0.71, 1.06, 33.33	0.71, 0.89, 0.62, 60.83
QGB	1.05, 0.91, 2.29, 94.63	1.73, 0.71, 3.99, 94.17	0.77, 0.88, 2.54, 100.0
BNN	1.18, 0.91, 0.78, 48.82	1.79, 0.73, 0.98, 40.0	0.73, 0.91, 0.78, 73.33
BNN + D	1.07, 0.91, 2.20, 96.75	1.90, 0.74, 3.20, 90.83	0.69, 0.90, 2.04, 100.0
Baseline MLP	1.09, 0.92, /, / ,	2.10, 0.80, /, / ,	0.77, 0.90, / , / ,
Baseline Frozen	1.18, 0.89, / , / ,	2.17, 0.58, / , / ,	0.81, 0.85, / , / ,
Baseline IRI 2016		3.39, 0.82, / , / ,	1.89, 0.92, / , / ,
VTEC:10°40°			
SE	1.83, 0.90, 0.92, 43.46	3.31, 0.80, 1.44, 41.67	1.32, 0.96, 0.66, 36.67
QGB	1.89, 0.89, 3.45, 94.17	3.35, 0.80, 4.59, 82.50	1.27, 0.96, 3.61, 100.0
BNN	1.95, 0.90, 1.20, 47.48	3.09, 0.85, 1.44, 39.17	1.40, 0.96, 1.24, 60.83
BNN + D	1.89, 0.90, 3.78, 95.11	2.94, 0.86, 4.24, 93.33	1.53, 0.94, 3.90, 99.17
Baseline MLP	1.92, 0.89, / , / ,	3.50, 0.85, / , / ,	1.48, 0.96, / , / ,
Baseline Frozen	2.22, 0.86, / , / ,	4.00, 0.72, / , / ,	1.33, 0.95, / , / ,
Baseline IRI 2016		5.63, 0.78, / , / ,	2.78, 0.95, / , / ,
VTEC:10°10°			
SE	2.08, 0.98, 1.32, 53.50	3.71, 0.96, 2.10, 39.12	2.19, 0.99, 1.22, 47.50
QGB	2.22, 0.98, 5.53, 96.21	3.98, 0.95, 6.51, 89.17	2.09, 0.99, 5.18, 95.83
BNN	2.28, 0.98, 1.66, 52.56	3.45, 0.96, 1.78, 40.00	1.90, 0.99. 1.60, 55.00
BNN + D	2.67, 0.97, 5.70, 97.38	3.63, 0.96, 7.02, 98.33	2.07, 0.99, 6.16, 100.0
Baseline MLP	2.34, 0.97, / , / ,	4.19, 0.96, / , / ,	2.16, 0.99, / , / ,
Baseline Frozen	2.40, 0.97, / , / ,	4.21, 0.94, / , / ,	2.31, 0.99, / , / ,
Baseline IRI 2016		8.41, 0.91, / , / ,	4.76, 0.94, / , / ,

Note. RMS stands for Root Mean Square, and Corr. for the correlation coefficient. RMS and Corr. are calculated between the median (QGB) or mean (SE, BNN, BNN + D) VTEC and ground truth. CI_{avg} represents the average distance of the lower and upper bounds from the forecast median (QGB) or mean (SE, BNN, BNN + D) VTEC. In(%) represents the percentage of ground truth within the 95% confidence intervals. The best results are highlighted in green. When all developed models have the same correlation coefficients, no values are highlighted, that is, all are black.

TECU for the high latitude point and 8 TECU for the low latitude point, while it is half lower during the quiet period in April 2017. The largest absolute GT VTEC values outside the CI occur during the strongest solar flare on 6 September, during moderate geomagnetic conditions on 7 September, during geomagnetic storms on 8 September, and at the beginning of the recovery period on 9 September. These results show that the forecast CI of the SE and BNN approaches exclude most of the sudden and intense VTEC variability during space weather events. However, the mean VTEC from the SE approach mostly achieves the lowest RMS for the entire test year. On the other hand, the QGB and BNN + D approaches provide 3 to 4 times wider CI, as shown in Figure 4, that contain more than 95% of GT in 2017 and even 100% during the quiet period. The largest absolute values of GT outside the CI are on 7 September, as well as during the first and second Dst minima, that is, the maximum intensity of the geomagnetic storm. The magnitude of the GT outside the CI is 4–5 TECU during the September 2017 space weather events and less than 2 TECU with much lower frequency during the quiet period. For instance, in the BNN + D approach, there is only one GT value outside the CI during the quiet period in April 2017. The median VTEC from the QGB approach mostly has a slightly lower correlation with GT than the other approaches, while the mean VTEC of the BNN + D approach has the highest correlation during intense space weather in Table 3. For the SE, BNN, and BNN + D approaches, the average width of the 90% CI from the VTEC mean

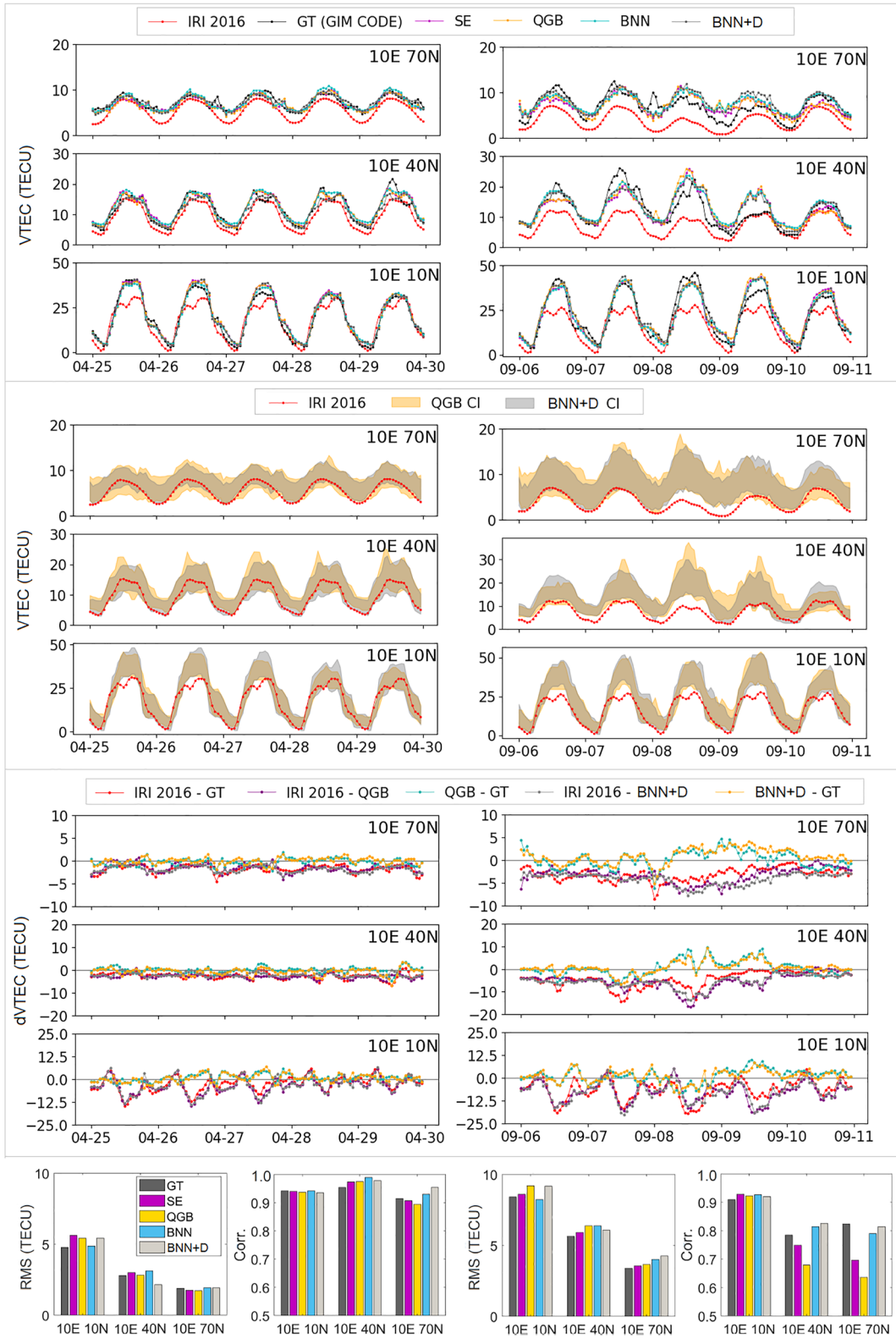
IRI 2016
GT (GIM CODE)
SE
QGB
BNN
BNN+D
10E 70N
10E 40N
10E 10N
VTEC (TECU)
04-25
04-26
04-27
04-28
04-29
04-30
09-06
09-07
09-08
09-09
09-10
09-11
IRI 2016
QGB CI
BNN+D CI
IRI 2016 - GT
IRI 2016 - QGB
QGB - GT
IRI 2016 - BNN+D
BNN+D - GT
dVTEC (TECU)
GT
SE
QGB
BNN
BNN+D
RMS (TECU)
Corr.
10E 10N
10E 40N
10E 70N

Figure 5.

can be estimated by multiplying the value for CI_{avg} in Table 3 by 0.8, and for the 99% CI by multiplying it by 1.3. Accordingly, the 90% CI will be 20% narrower, and the 99% CI will be 30% wider than the 95% CI.

The mean/median VTEC of the studied approaches mostly outperforms the baseline frozen ionosphere, with the most significant improvement from around 20%–30% during severe space weather (6–10 September) for all three investigated VTEC grid points. Their differences are not as significant for the quiet period in April 2017, that is, from 0.1 to 0.3 RMS. The mean/median VTEC values of the developed models have lower RMS than the baseline MLP model for most of the study cases. In particular, for the storm period, they improve the RMS by about 0.50 TECU or 15%, while the MLP model still maintains high correlations with GT. The RMS of the IRI 2016 is about twice as large as the RMS for the mean/median VTEC of the developed ML-based UQ VTEC models for both quiet and storm study cases.

The detailed analysis for the IRI 2016 is presented in Figure 5. In April 2017, the IRI is mainly within the CI of QGB or BNN + D, while in September 2017, it is sometimes at the edge and sometimes outside these intervals. The differences between the IRI VTEC values and the GT VTEC, and between IRI and the median VTEC of QGB and mean VTEC of BNN + D models are similar. They are mostly more prominent than the differences between the median/mean VTEC of QGB/BNN + D and the GT data. The enormous IRI differences exist for the low-latitude VTEC position, where VTEC from IRI is underestimated by up to more than 20 TECU. Vertical Total Electron Content from IRI agrees much better with the mean/median VTEC values from the ML-based UQ VTEC models and the GT data in April 2017 than September 2017. Consequently, the RMS values between IRI and GT and ML-based UQ VTEC models are smaller in April 2017, that is, when the ionosphere is quiet, while they are twice as significant when the ionosphere is disturbed in September 2017. This space weather effect is also reflected in the correlation coefficients between IRI and GT, as well as ML-based UQ VTEC models.

The upper and LB of the 95% CI estimated using different approaches are visualized in Figure 6 without the VTEC mean/median, that is, they are adjusted around $y = 0$. In the case of the quiet period (Figure 6, left), the QGB and BNN + D CI are similar in size for the mid-latitude grid point, while the QGB CI is wider for the high-latitude point, and the BNN + D confidence upper bound is slightly larger for the low-latitude point. The SE and BNN CI are of a similar order of magnitude. The main difference is that the BNN CI is smoother and more constant over the study period, while the SE CI is variable. During the storm case (Figure 6, right), the CI become wider as the changes in the GMF occur. The SE and QGB CI are about two times wider and more variable on the day of the geomagnetic storm maximum (8 September) and the following day of the recovery phase (9 September), while the BNN and BNN + D CI slightly increase. The largest upper confidence bound for high- and mid-latitude points in this period comes from the QGB approach, while for the low-latitude point, the QGB and BNN + D upper confidence bounds are similar in size. For both study cases, it can be seen that the QGB and SE CI are more variable and have frequent peaks, while for BNN and BNN + D, they are smoother and more consistent from day-to-day. The CI of all approaches are wider around local noon for the mid-latitude point, while for the low-latitude point, an additional increase in the upper bound is visible after sunset and lasts for several hours. Post-sunset increase in the QGB and BNN + D upper low-latitude VTEC bounds is visible for 6 to 9 September with F10.7 > 110 sfu, and from 25 to 29 April during a period of low geomagnetic activity, with both periods close to equinox. The effect is more pronounced in QGB. The post-sunset VTEC enhancement has been detected at low latitudes within the equatorial ionization anomaly using actual VTEC observations in Dashora et al. (2019), Kutiev et al. (2007), Kumar et al. (2022), J. Liu et al. (2020). It develops 2–3 hr after sunset, with a peak around 7:00–8:00 p.m. local time (Kumar et al., 2022; Kutiev et al., 2007), and occurs during prolonged periods of low geomagnetic activity (Kutiev et al., 2007), as well as during geomagnetic storms (Dashora et al., 2019), with stronger intensity around equinoxes (J. Liu et al., 2020), and when the F10.7 solar flux exceeds 110 sfu (Kumar et al., 2022). Therefore, the patterns of increase in the upper low-latitude VTEC bounds after sunset are consistent with observations of the low-latitude VTEC post-sunset enhancement reported in previous studies.

The results of the analysis in Figure 6 show that the CI exhibit variations depending on daytime/nighttime, solar irradiance, space weather conditions, that is, geomagnetic storms, and the post-sunset ionosphere enhancement at

Figure 5. First panel: Vertical Total Electron Content (VTEC) from International Reference Ionosphere (IRI) 2016, ground truth (GT) data, and the mean/median VTEC values from uncertainty quantification ML-based VTEC models. Second panel: IRI 2016 and confidence intervals of Quantile Gradient Boosting (QGB) and BNN + D. Third panel: VTEC differences between IRI 2016 and GT, the QGB median VTEC, and the BNN + D mean VTEC, as well as the differences between QGB/BNN + D and GT. Fourth panel: RMS and correlation of GT and ML-based mean/median VTEC values with respect to IRI 2016. Left: 25–29 April 2017, right: 6–10 September 2017.

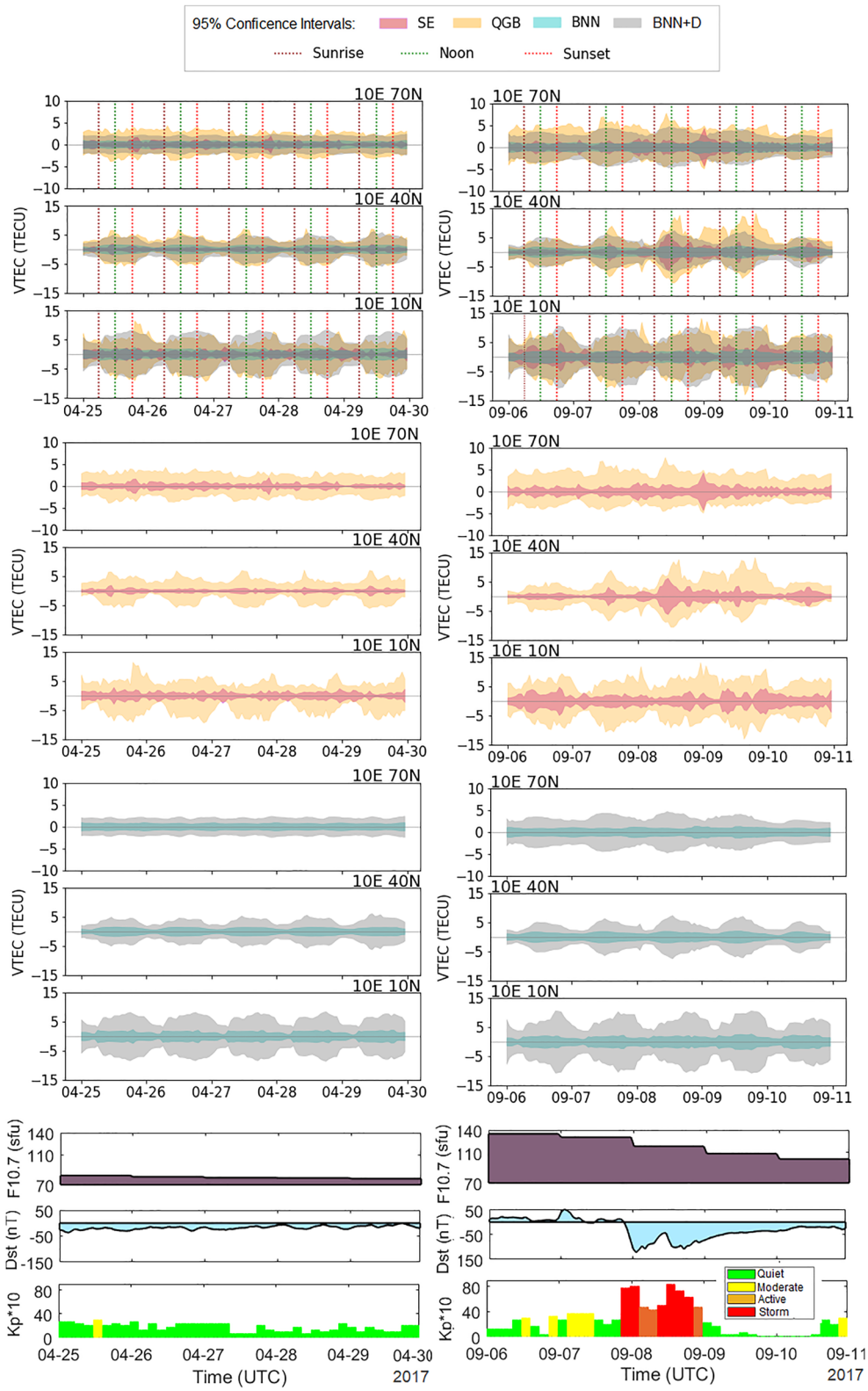

Figure 6. 95% confidence interval of all developed uncertainty quantification ML-based Vertical Total Electron Content models (first panel), SE and Quantile Gradient Boosting (second panel), Bayesian neural network (BNN) and BNN + D (third panel) for three selected grid points. Fourth panel: indices of F10.7, Dst, and Kp · 10. Left: 25–29 April 2017; right: 6–10 September 2017.

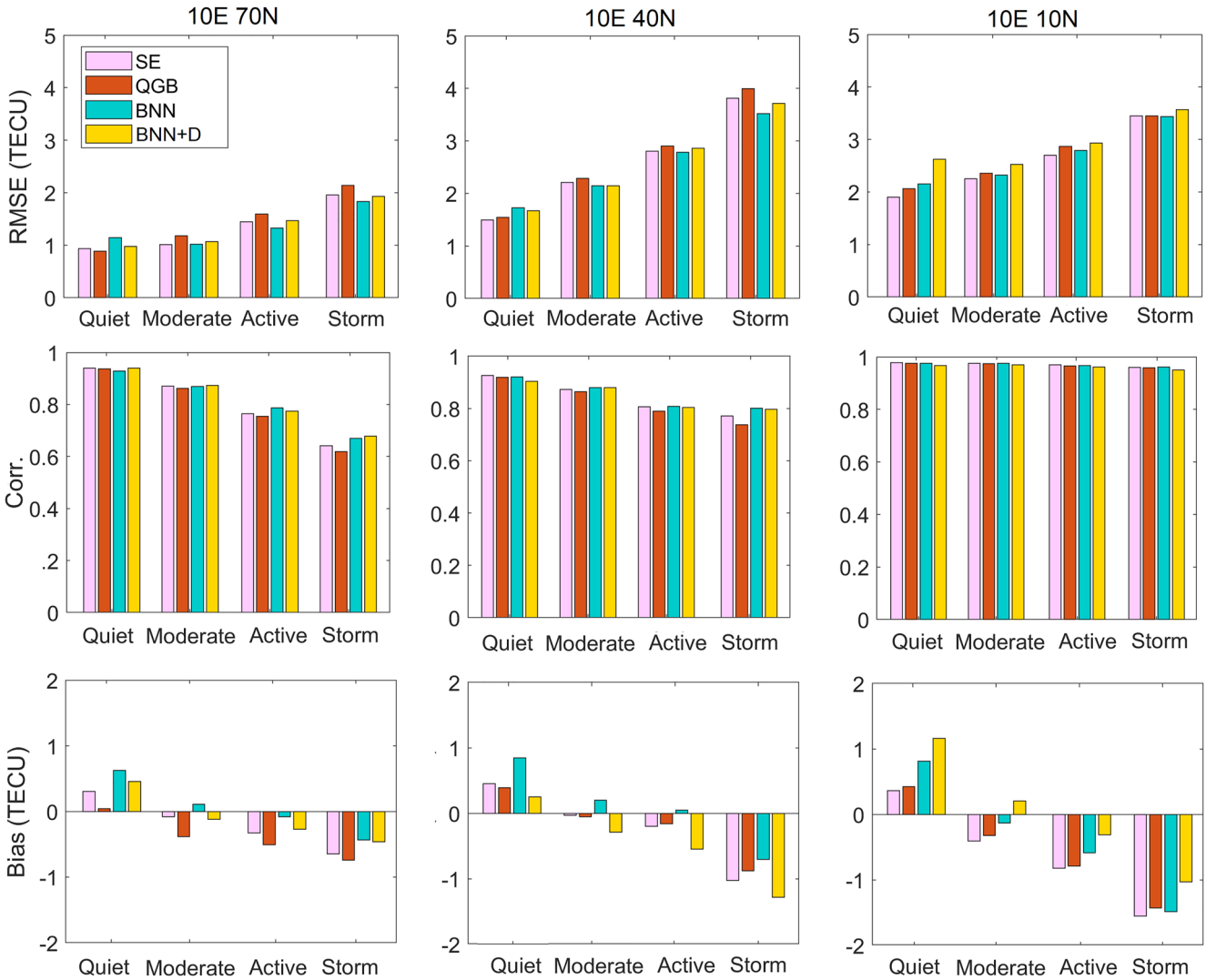

Figure 7. Statistics of mean/median Vertical Total Electron Content forecast from the developed models to ground truth versus Kp index for 2017. Top: RMS, mid: correlation coefficients (Corr.), bottom: bias. The labels quiet, moderate, active and storm correspond to $Kp < 3$, $3 \leq Kp < 4$, $4 \leq Kp < 5$, and $Kp \geq 5$, respectively.

low latitudes. Therefore, they are narrower during the night, wider around local noon for the mid-latitude point, wider and more variable with the change of Kp and Dst indices, and wider after sunset for the low latitude point under certain conditions mentioned above.

Further analysis is performed regarding geomagnetic activity in Figure 7. The forecast mean/median VTEC accuracy in terms of RMS and correlation coefficients decreases with increasing geomagnetic activity. The biases are largest and negative during storms, suggesting that the models underestimate the mean/median VTEC for storms. Due to the complex, distinct VTEC irregularities during different geomagnetic storms, the lack of VTEC samples covering different geomagnetic storms under different dependent factors such as storm intensity, season, magnetic local time, storm onset time, magnetic latitude and solar cycle phase (Greer et al., 2017; J. Liu et al., 2010; Vijaya Lekshmi et al., 2011), as well as the overall presence of storm events in the data set, resulting in a high imbalance compared to the quiet condition samples (see Figure 7 in Natras et al. (2022a)), the developed ML-based UQ VTEC models have lower accuracy in forecasting the mean/median VTEC during storms. To evaluate the full performance of the ML-based UQ VTEC models and achieve realistic accuracy representation, we need to consider the full probabilistic prediction, that is, the CI.

Figure 8 represents boxplots of the GT VTEC outside the forecast 95% CI concerning different geomagnetic activity levels. The analysis is performed only for data samples where the GT falls outside the forecast CI, which

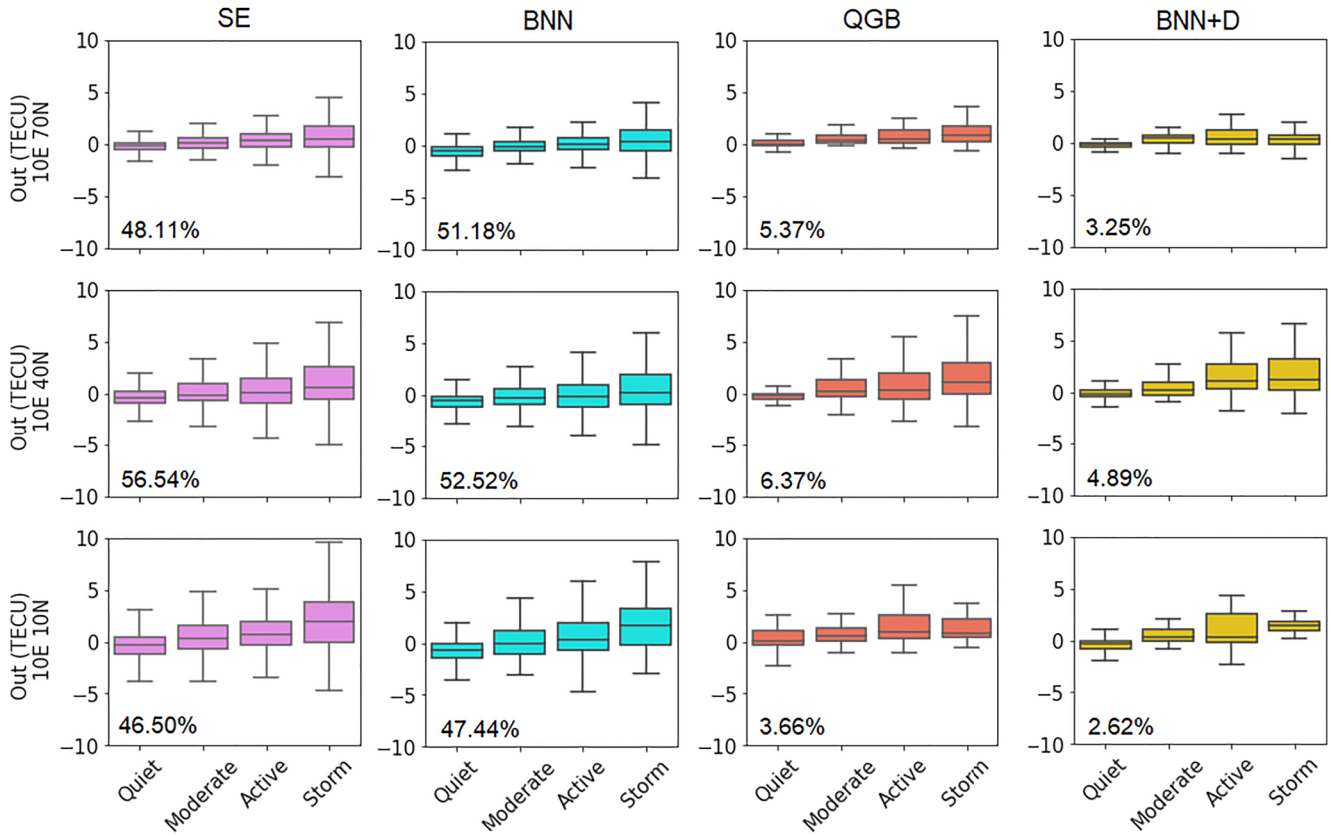

Figure 8. Boxplots for ground truth (GT) outside the forecast confidence intervals (CI) versus different geomagnetic conditions (quiet, moderate, active, and storm correspond to $Kp < 3$, $3 \leq Kp < 4$, $4 \leq Kp < 5$, and $Kp \geq 5$, respectively). In each graph, the percentage of outliers in 2017 is given (bottom left). It corresponds to the data samples for which GT falls outside the forecast CI. Positive value: the amount by which GT is higher than the upper CI bound; negative value: the amount by which GT is lower than the lower CI bound. The boxes (the interquartile range) represent the range between the 25th (first quartile) and 75th percentile (third quartile), that is, the middle 50%. The gray line in each box corresponds to the median. The gray lines outside the boxes represent the lower 25% and the upper 25% of the values, with the ends of the line representing the minimum and maximum values. First row: 10°70°, mid row: 10°40°, third column: 10°10°. From left to right: SE, Quantile Gradient Boosting, Bayesian neural network (BNN), BNN + D.

we can refer to here as outliers from the CI. Most outliers are between 0 and 1 TECU outside of the CI. There is a clear tendency for the interquartile range and the maximum absolute values of the outliers to increase with increasing geomagnetic activity in SE and BNN models. The interquartile range and the maximum and minimum outliers also tend to be the largest in these models. In contrast, the BNN + D and QGB models have the lowest percentage of outliers: 3%–5% and 4%–6%, respectively. Considering that the CI are set at 95%, outliers up to 5% from the CI indicate reliable performance. According to these results, both BNN + D and QGB approaches achieve the target of 95% confidence. Moreover, the amount of outliers from the CI for these two methods is less affected by geomagnetic conditions.

The relative importance of input features for probabilistic VTEC forecasting by the QGB model (Figure 9) is estimated for the upper confidence bound (top), median VTEC (middle), and lower confidence bound (bottom) using the methodology presented in Text S1 in Supporting Information S1. For the median VTEC, the most important input feature is the lagged VTEC at time step t_i for forecasting VTEC at time step t_{i+24h}. This is due to the prevailing diurnal VTEC variations, where day-to-day VTEC usually does not change much during quiet conditions. On the other hand, other input features have much greater importance in estimating the lower and upper limits, such as the AE, Kp, Dst, and SW indices. Here, the objective function minimizes the positive and negative residuals between the GT and the model results for the upper and LB, respectively; see Equation 7. These residuals are more strongly influenced by solar and geomagnetic activity than the median VTEC. Thus, the lagged VTEC contributes 20%–50% less to the confidence bounds estimate than to the median VTEC estimate, while the space weather input features increase their contribution. These results suggest that the CI are determined by the space weather features in addition to the VTEC-related features.

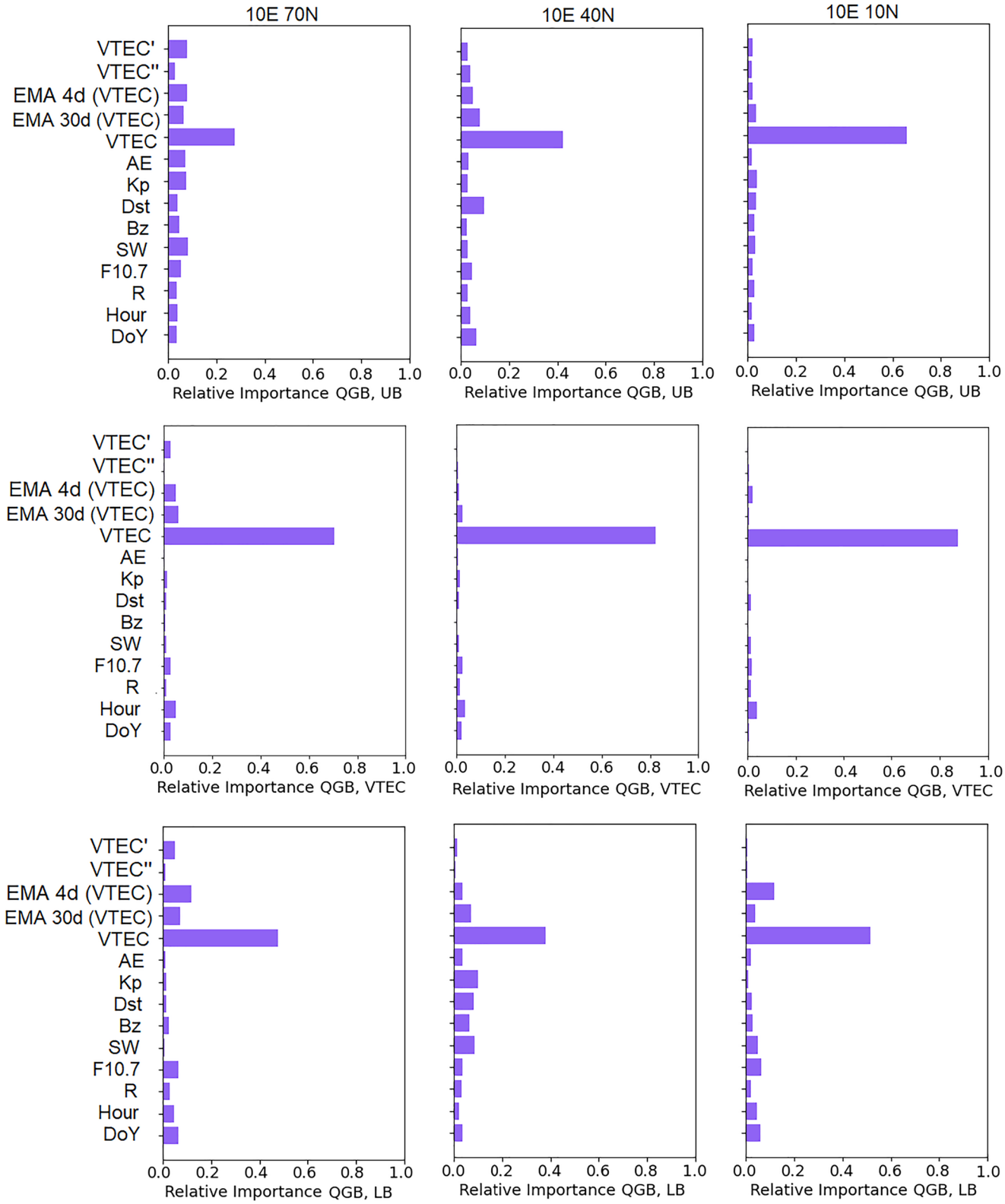

Figure 9. The relative importance of input features for the 1-day Quantile Gradient Boosting Vertical Total Electron Content (VTEC) forecast, consisting of upper bound (top), median VTEC (mid), and lower bound (bottom). VTEC' and VTEC'' represent the first and second derivatives, respectively; exponential moving average (EMA) 4d (VTEC) and EMA 30d (VTEC) represent EMAs of VTEC over 4 and 30 days, respectively. The input features refer to time step i, while the output or forecast is the VTEC at time step $i + 24$, that is, $y_i = \text{VTEC}_{(i+24)}$. Left column: 10° 70°, mid column: 10° 40°, right column: 10° 10°.

Table 4
Computational Cost for a Single Training Run (Mid) and a Test Run (Left) for a Single Vertical Total Electron Content Grid Point Using the NVIDIA Tesla P100 GPU With 16 GB Memory

Model	Training	Testing
SE	1,275 s (~20 min)	1.35 s
QGB	50–80 s (~1 min)	<0.1 s
BNN	1,917–3,648 s (~30–60 min)	1.37 s
BNN + D	5,900 s (~100 min)	1.40 s

As for the computational complexity analysis in Table 4, the two BNN-based approaches are the most computationally intensive. A single training iteration with two years of data takes about 1 hr or more on the NVIDIA Tesla P100 GPU computing processor with 16 GB, which can be considered a disadvantage of the BNN method. The most computationally efficient model is QGB, which takes only 1 min for a training iteration with 2 years of data. When the models are trained and optimized, the execution is fast and takes 1 s for 1 year of data.

4. Conclusion

This work is the first to thoroughly examine probabilistic VTEC forecasting using ML techniques and quantifying uncertainties. In addition to forecasting a single VTEC value, the models estimate 95% CI to provide information on how confident and reliable results are by considering the uncertainties in the model parameters and/or data. In summary, we have implemented and analyzed several approaches for 1-day UQ ML-based VTEC forecasting, including:

- SE of multiple models trained with different tree-based learning algorithms and data sets to estimate uncertainties as ensemble spread,
- QGB, in which probabilistic output is estimated by minimizing quantile loss, with quantiles set at 0.025 and 0.975, for the lower and upper confidence bound, respectively, and 0.50 for median VTEC to capture the data uncertainties,
- Bayesian Neural Network (BNN), where the probability distributions of the parameters are learned to estimate the model uncertainty,
- BNN including data uncertainty (BNN + D) to capture the data uncertainty.

The findings can be summarized as follows:

1. The SE and BNN approaches provide the lowest uncertainties and, thus, overconfident results. In reality, the GT VTEC in 2017 is outside the forecast CI about 50% of the time.
2. The approaches that capture data uncertainties, QGB and BNN + D, provide wider CI that contain GT around 95% of the time and are, therefore, more realistic and reliable.
3. As for the forecasting of the mean/median VTEC, the SE approach often yields the lowest RMS value, demonstrating the power of an ensemble to improve the accuracy of the deterministic estimate. On the other hand, BNN tends to provide the highest correlations to GT, especially during the storm.
4. The relative importance of the input features shows that the CI for the QGB model are determined by space weather indices in addition to VTEC-related input features, with lagged VTEC dominating.
5. CI, especially of QGB, exhibit variations depending on the daytime/nightime, solar irradiance, geomagnetic activity, and post-sunset low-latitude ionosphere enhancement.
6. The most computationally intensive method is BNN + D, while QGB is the fastest.
7. The data uncertainties are at least three times larger than the model parameter uncertainties.

The advantages and disadvantages of each investigated UQ method for VTEC forecasting are outlined in Table 5.

Based on these findings, the probabilistic VTEC forecasting that only considers the model parameter uncertainties are insufficient. An ML-based model trained with different learning algorithms using the same/similar data sets performs similarly because it learns an approximation function from similar data, resulting in smaller discrepancies between the solutions of different ML-based models in an ensemble. The ensemble approach for UQ could be improved by training the base models on different subsets of data covering different study cases, which would increase diversity and randomness among ensemble members and may better describe uncertainties. Probabilistic VTEC modeling and forecasting, which accounts for both model parameters and data uncertainties, would be the optimal solution, as shown by the BNN + D results. Due to the computational complexity of the BNN + D approach, modification may be required to obtain a computationally efficient and accurate model. In this context, the advantage of fast gradient boosting computation on decision trees can be exploited. The QGB model could be improved by adding the model uncertainties, for example, via an ensemble of multiple models (with data uncertainty-informed base models) or virtual ensembles (Malinin et al., 2021) using a single gradient boosting model. Instead of estimating multiple quantile functions separately, the method can be modified to

Table 5
Advantages and Disadvantages of Different Investigated Uncertainty Quantification Approaches for Vertical Total Electron Content Forecasting

	SE	QGB	BNN	BNN + D
PROS	Improved mean VTEC No distribution assumption	Fast to train CI ~ 95% GT No distribution assumption	Higher Corr. to GT	CI > 95% GT
CONS	Many models to train Uncertainty too small	Estimate each quantile	Slow to train Gaussian distribution Uncertainty too small	Slow to train Gaussian distribution

Note. CI stands for confidence interval, GT for ground truth, and Corr. for correlation coefficients.

estimate them simultaneously (e.g., X. Han et al., 2021; Y. Liu & Wu, 2011). Moreover, adding information about the uncertainty of the input data directly into a model can further improve the probabilistic estimation of output and provide a more realistic representation of the uncertainties (e.g., Kiani Shahvandi & Soja, 2022). It is also important to note that we assumed GIM CODE data to be GT, which is not error-free. In further work, GT uncertainty information may also be included, for example, as an additional input value to the model as in Kiani Shahvandi and Soja (2022).

The results from this study show that the uncertainty arising from the data is much larger than that of the model parameters. Therefore, the input data of an ML-based ionosphere model are much more important to be considered for future improvements. Further steps may include investigating and incorporating new input observations, extracting new input features for VTEC modeling and forecasting that can characterize the effects of space weather on the ionosphere in a way that is more helpful to the learning process. Some input observations, such as the F10.7 and Kp indexes, have lower resolution. Including data with higher temporal resolution and minimizing the need to interpolate values may also reduce uncertainties.

As can be seen from the results, the uncertainties during the space weather event in September 2017 are up to 1.5 to 2 times larger than during the quiet period in April 2017. The ionospheric response to a geomagnetic storm depends on several factors that lead to distinct VTEC irregularities during different storms, as well as on the overall presence of storm events in the data set. The VTEC response to geomagnetic disturbances depends not only on the intensity of the storm, but also on the season, magnetic local time, storm onset time, magnetic latitude, and solar cycle phase (Greer et al., 2017; J. Liu et al., 2010; Vijaya Lekshmi et al., 2011). Therefore, it varies from one storm to another, making it difficult for a learning algorithm to find an approximation function that generalizes to all storms. Another challenge is the small number of storm samples in the training data. The analysis by Natras et al. (2022a) shows that only around 11% of the samples from January 2015 to December 2016 belong to geomagnetic active and storm conditions, even though these years contain the highest number of geomagnetic storms in solar cycle 24. If the training data set contained balanced instances of quiet and storm periods, the forecast accuracy during a space weather event could be improved and the associated uncertainties reduced.

Recommended solutions for the imbalanced data set to be explored in future work include improving the input features for learning rare space weather-related VTEC signatures, training on the balanced data set achieved with oversampling or undersampling, or developing a cost-sensitive solution that can adjust the penalty for the degree of importance assigned to the minority case. Another possible solution is combining physical laws and equations with ML to develop a physically informed ML-based VTEC model, which could improve space weather modeling when only few training examples of space weather events exist and reduce uncertainties.

Since dynamic solar-terrestrial processes and space weather govern the ionosphere, and the VTEC quantity is essential for positioning applications and early-warning systems of space weather effects, it is crucial to include reliability and confidence information in VTEC and space weather forecasting. Moreover, such information will increase the explain ability and interpretability of ML-based ionosphere modeling and forecasting, and trust in ML results in general. Therefore, we encourage further work on uncertainty estimation to produce trustworthy probabilistic ionosphere and space weather forecasts. We hope that the research community will begin to incorporate probabilistic frameworks into their ML solutions alongside the tremendous amount of work exploring

various learning algorithms for VTEC approximation. This study is a starting point for discussing and integrating UQ solutions into ML-based VTEC forecasting and will hopefully lead to further ML-based ionosphere and space weather studies that take uncertainties into account.

Data Availability Statement

Software used to implement machine learning (ML) approaches are ScikitLearn (Pedregosa et al., 2011) and TensorFlow (Abadi et al., 2015). The figures were created in Python using Seaborn (Waskom, 2021) and Matplotlib (Hunter, 2007), and in Matlab (MATLAB, 2020). Global ionosphere maps (GIM) produced by the Center for Orbit Determination in Europe (CODE) at the University of Bern, available in Dach et al. (2020), were used to prepare the VTEC data in this study. Other input data to the ML-based VTEC models: sunspot number, F10.7 solar radio flux, solar wind plasma speed, Bz index, Dst index, Kp index, and AE index are publicly available via NASA/GSFC's OMNIWeb (King & Papitashvili, 2005). The IRI 2016 was retrieved from the Community Coordinated Modeling Center (CCMC) Instant-Run System of NASA Goddard Space Flight Center at https://kauai.ccmc.gsfc.nasa.gov/instantrun/iri. The dataset containing the probabilistic VTEC forecast results for the year 2017 from the four ML uncertainty quantification approaches presented and discussed in this study is openly available under the Creative Commons Attribution 4.0 International license at Zenodo (Natras et al., 2023b). The codes defining the architecture of the BNN and BNN + D VTEC models, the model development process using training and cross-validation data, and their evaluation using test data can be found in Natras (2023a). The codes for loading the QGB VTEC models and evaluating them using test data are provided along with the developed QGB VTEC models in Natras (2023b).

Acknowledgments
This research was funded by Research Grants—Doctoral Programmes in Germany from the German Academic Exchange Service (in German: Deutscher Akademischer Austauschdienst, DAAD). Open access funding enabled and organized by Projekt DEAL. The authors acknowledge the use of GIM products of CODE from University of Bern; the OMNIWeb CDAWeb service of NASA/GSFC's Space Physics Data Facility and OMNI data; CCMC Instant-Run System of NASA Goddard Space Flight Center and IRI 2016 data. Open Access funding enabled and organized by Projekt DEAL.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Retrieved from https://www.tensorflow.org/

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Amell, A., Eriksson, P., & Pfreundschuh, S. (2022). Ice water path retrievals from meteosat-9 using quantile regression neural networks. *Atmospheric Measurement Techniques*, *15*(19), 5701–5717. https://doi.org/10.5194/amt-15-5701-2022

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural network. In *Proceedings of the 32nd international conference on machine learning* (Vol. 37, pp. 1613–1622).

Bottou, L. (2012). Stochastic gradient descent tricks. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural networks: Tricks of the trade* (2nd ed., pp. 421–436). Springer. https://doi.org/10.1007/978-3-642-35289-8_25

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chan, J. S. K. (2021). Predicting loss reserves using quantile regression running title: Quantile regression loss reserve models. *Journal of Data Science*, *13*(1), 127–156. https://doi.org/10.6339/JDS.201501_13(1).0008

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794).

Dach, R., Schaer, S., Arnold, D., Kalarus, M. S., Prange, L., Stebler, P., et al. (2020). Code final product series for the IGS [Dataset]. Astronomical Institute, University of Bern. Retrieved from http://www.aiub.unibe.ch/download/CODE

Dashora, N., Suresh, S., & Niranjan, K. (2019). Interhemispheric asymmetry in response of low-latitude ionosphere to perturbation electric fields in the main phase of geomagnetic storms. *Journal of Geophysical Research: Space Physics*, *124*(8), 7256–7282. https://doi.org/10.1029/2019JA026671

Duan, T., Anand, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., & Schuler, A. (2020). Ngboost: Natural gradient boosting for probabilistic prediction. In *International conference on machine learning* (pp. 2690–2700).

Ferreira, A. A., Borges, R. A., Paparini, C., Ciraolo, L., & Radicella, S. M. (2017). Short-term estimation of GNSS TEC using a neural network model in Brazil. (Studies on mesosphere, thermosphere and ionosphere from equatorial to mid latitudes—Recent investigations and improvements—Part 1). *Advances in Space Research*, *60*(8), 1765–1776. https://doi.org/10.1016/j.asr.2017.06.001

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Greer, K. R., Immel, T., & Ridley, A. (2017). On the variation in the ionospheric response to geomagnetic storms with time of onset. *Journal of Geophysical Research: Space Physics*, *122*(4), 4512–4525. https://doi.org/10.1002/2016JA023457

Han, X., Dasgupta, S., & Ghosh, J. (2021). Simultaneously reconciled quantile forecasting of hierarchically related time series. In A. Banerjee & K. Fukumizu (Eds.), *Proceedings of the 24th international conference on artificial intelligence and statistics* (Vol. 130, pp. 190–198). PMLR. Retrieved from https://proceedings.mlr.press/v130/han21a.html

Han, Y., Wang, L., Fu, W., Zhou, H., Li, T., & Chen, R. (2022). Machine learning-based short-term gps tec forecasting during high solar activity and magnetic storm periods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 115–126. https://doi.org/10.1109/JSTARS.2021.3132049

Hu, A., Shneider, C., Tiwari, A., & Camporeale, E. (2022). Probabilistic prediction of dst storms one-day-ahead using full-disk soho images. *Space Weather*, *20*(8), e2022SW003064. https://doi.org/10.1029/2022SW003064

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Imtiaz, N., Younas, W., & Khan, M. (2020). Response of the low- to mid-latitude ionosphere to the geomagnetic storm of September 2017. *Annales Geophysicae*, *38*(2), 359–372. https://doi.org/10.5194/angeo-38-359-2020

Kaselimi, M., Voulodimos, A., Doulamis, N., Doulamis, A., & Delikaraoglou, D. (2022). Deep recurrent neural networks for ionospheric variations estimation using GNSS measurements. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–15. https://doi.org/10.1109/TGRS.2021.3090856

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of the 31st international conference on neural information processing systems* (Vol. 30, pp. 5580–5590).

Kiani Shahvandi, M., & Soja, B. (2022). Inclusion of data uncertainty in machine learning and its application in geodetic data science, with case studies for the prediction of earth orientation parameters and GNSS station coordinate time series. *Advances in Space Research*, *70*(3), 563–575. https://doi.org/10.1016/j.asr.2022.05.042

King, J. H., & Papitashvili, N. E. (2005). Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data. *Journal of Geophysical Research*, *110*(A2), A02104. https://doi.org/10.1029/2004JA010649

Koch, K. R. (2018). Bayesian statistics and Monte Carlo methods. *Journal of Geodetic Science*, *8*(1), 18–29. https://doi.org/10.1515/jogs-2018-0003

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *The Journal of Economic Perspectives*, *15*(4), 143–156. https://doi.org/10.1257/jep.15.4.143

Kumar, A., Chakrabarty, D., Pandey, K., & Yadav, A. K. (2022). Solar flux dependence of post-sunset enhancement in vertical total electron content over the crest region of equatorial ionization anomaly. *Journal of Geophysical Research: Space Physics*, *127*(5), e2021JA030156. https://doi.org/10.1029/2021JA030156

Kutiev, I., Otsuka, Y., Saito, A., & Tsugawa, T. (2007). Low-latitude total electron content enhancement at low geomagnetic activity observed over Japan. *Journal of Geophysical Research*, *112*(A7), 893. https://doi.org/10.1029/2007JA012385

Lee, S., Ji, E.-Y., Moon, Y.-J., & Park, E. (2020). One-day forecasting of global TEC using a novel deep learning model. *Space Weather*, *19*(1), 2020SW002600. https://doi.org/10.1029/2020SW002600

Licata, R. J., & Mehta, P. M. (2022). Uncertainty quantification techniques for data-driven space weather modeling: Thermospheric density application. *Scientific Reports*, *12*(1), 1–17. https://doi.org/10.1038/s41598-022-11049-3

Liu, J., Zhang, D., Mo, X., Xiong, C., Hao, Y., & Xiao, Z. (2020). Morphological differences of the northern equatorial ionization anomaly between the eastern Asian and American sectors. *Journal of Geophysical Research: Space Physics*, *125*(3), e2019JA027506. https://doi.org/10.1029/2019JA027506

Liu, J., Zhao, B., & Liu, L. (2010). Time delay and duration of ionospheric total electron content responses to geomagnetic disturbances. *Annales Geophysicae*, *28*(3), 795–805. https://doi.org/10.5194/angeo-28-795-2010

Liu, L., Morton, Y. J., & Liu, Y. (2022). Ml prediction of global ionospheric TEC maps. *Space Weather*, *20*(9), e2022SW003135. https://doi.org/10.1029/2022SW003135

Liu, L., Zou, S., Yao, Y., & Wang, Z. (2020). Forecasting global ionospheric TEC using deep learning approach. *Space Weather*, *18*(11), e2020SW002501. https://doi.org/10.1029/2020SW002501

Liu, Y., & Wu, Y. (2011). Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of Nonparametric Statistics*, *23*(2), 415–437. https://doi.org/10.1080/10485252.2010.537

Malinin, A., Prokhorenkova, L., & Ustimenko, A. (2021). Uncertainty in gradient boosting via ensembles. In *International conference on learning representations*. Retrieved from https://openreview.net/forum?id=1Jv6b0Zq3qi

MATLAB. (2020). *Version 2020a*. The MathWorks Inc.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Natras, R. (2023a). Randa-lab/Bayesian_Neural_Network_Probabilistic_Io nosphere_VTEC: Bayesian_Neural_Network_Probabilistic_Ionosphere. *Zenodo*. https://doi.org/10.5281/zenodo.7858906

Natras, R. (2023b). Randa-lab/Quantile_Gradient_Boosting_for_Probabili stic_VTEC: Quantile_Gradient_Boosting_Probabilisti c_Ionosphere_Evaluation. *Zenodo*. https://doi.org/10.5281/zenodo.7858661

Natras, R., Goss, A., Halilovic, D., Magnet, N., Mulic, M., Schmidt, M., & Weber, R. (2023a). Regional ionosphere delay models based on cors data and machine learning. *NAVIGATION: Journal of the Institute of Navigation*, *70*(3), navi.577. https://doi.org/10.33012/navi.577

Natras, R., Soja, B., & Schmidt, M. (2022a). Ensemble machine learning of Random Forest, AdaBoost and XGBoost for vertical total electron content forecasting. *Remote Sensing*, *14*(15), 3547. https://doi.org/10.3390/rs14153547

Natras, R., Soja, B., & Schmidt, M. (2022b). Machine learning ensemble approach for ionosphere and space weather forecasting with uncertainty quantification. In *2022 3rd URSI Atlantic and Asia Pacific Radio Science Meeting (AT-AP-RASC)* (pp. 1–4). https://doi.org/10.23919/AT-AP-RASC54737.2022.9814334

Natras, R., Soja, B., & Schmidt, M. (2023b). Dataset of machine learning forecasted VTEC from paper: Uncertainty quantification for machine learning-based ionosphere and space weather forecasting. *Zenodo*. https://doi.org/10.5281/zenodo.7741342

Orus Perez, R. (2019). Using tensorflow-based neural network to estimate GNSS single frequency ionospheric delay (iononet). *Advances in Space Research*, *63*(5), 1607–1618. https://doi.org/10.1016/j.asr.2018.11.011

Özkan, A. (2022). An artificial neural network model in predicting VTEC over central Anatolia in Turkey. *Geodesy and Geodynamics*, *14*(2), 130–142. https://doi.org/10.1016/j.geog.2022.07.004

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rahaman, R., & Thiery, A. (2021). Uncertainty quantification and deep ensembles. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 20063–20075). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2021/file/a70dc40477bc2adceef4d2c90f47eb82-Paper.pdf

Schaer, S. (1999). *Mapping and predicting the earth's ionosphere using the global positioning system* (Vol. 59). Institut für Geodäsie und Photogrammetrie, Eidg. Technische Hochschule.

Siddique, T., Mahmud, M. S., Keesee, A. M., Ngwira, C. M., & Connor, H. (2022). A survey of uncertainty quantification in machine learning for space weather prediction. *Geosciences*, *12*(1), 27. https://doi.org/10.3390/geosciences12010027

Tagasovska, N., & Lopez-Paz, D. (2019). Single-model uncertainties for deep learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2019/file/73c03186765e199c116224b68adc5fa0-Paper.pdf

Vasseur, S., & Aznarte, J. (2021). Comparing quantile regression methods for probabilistic forecasting of NO_2 pollution levels. *Scientific Reports*, *11*(1), 11592. https://doi.org/10.1038/s41598-021-90063-3

Vijaya Lekshmi, D., Balan, N., Tulasi Ram, S., & Liu, J. Y. (2011). Statistics of geomagnetic storms and ionospheric storms at low and mid latitudes in two solar cycles. *Journal of Geophysical Research*, *116*(A11), 530. https://doi.org/10.1029/2011JA017042

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021