Journal of Geodesy (2024) 98:23
https://doi.org/10.1007/s00190-024-01829-2

ORIGINAL ARTICLE

Global, spatially explicit modelling of zenith wet delay with XGBoost

Laura Crocetti[1] · Matthias Schartner[1] · Florian Zus[2] · Wenyuan Zhang[1,3] · Gregor Moeller[1,4] · Vicente Navarro[5] · Linda See[6] · Konrad Schindler[1] · Benedikt Soja[1]

Received: 22 December 2022 / Accepted: 19 February 2024

Abstract

Radio signals transmitted by Global Navigation Satellite System (GNSS) satellites experience tropospheric delays. While the hydrostatic part, referred to as zenith hydrostatic delay (ZHD) when mapped to the zenith direction, can be analytically modelled with sufficient accuracy, the wet part, referred to as zenith wet delay (ZWD), is much more difficult to determine and needs to be estimated. Thus, there exist several ZWD models which are used for various applications such as positioning and climate research. In this study, we present a data-driven, global model of the spatial ZWD field, based on the Extreme Gradient Boosting (XGBoost). The model takes the geographical location, the time, and a number of meteorological variables (in particular, specific humidity at several pressure levels) as input, and can predict ZWD anywhere on Earth as long as the input features are available. It was trained on ZWDs at 10718 GNSS stations and tested on ZWDs at 2684 GNSS stations for the year 2019. Across all test stations and all observations, the trained model achieved a mean absolute error of 6.1 mm, respectively, a root mean squared error of 8.1 mm. Comparisons of the XGBoost-based ZWD predictions with independently computed ZWDs and baseline models underline the good performance of the proposed model. Moreover, we analysed regional and monthly models, as well as the seasonal behaviour of the ZWD predictions in different climate zones, and found that the global model exhibits a high predictive skill in all regions and across all months of the year.

Keywords Zenith wet delay (ZWD) · Global predictions · XGBoost · Machine learning (ML) · GNSS

1 Introduction

Radio signals emitted by Global Navigation Satellite System (GNSS) satellites propagate through the atmosphere before being received on Earth. The signals are delayed and bent when travelling through the neutral atmosphere and this delay can be estimated (Nilsson et al. 2013). Typically, the tropospheric delays in zenith direction, i.e. the delays in the lowest

✉ Laura Crocetti
lcrocetti@ethz.ch

Matthias Schartner
mschartner@ethz.ch

Florian Zus
zusflo@gfz-potsdam.de

Wenyuan Zhang
zhangwy@cumt.edu.cn

Gregor Moeller
gregor.moeller@geo.tuwien.ac.at

Vicente Navarro
vicente.navarro@esa.int

Linda See
see@iiasa.ac.at

Konrad Schindler
schindler@ethz.ch

Benedikt Soja
benedikt.soja@geod.baug.ethz.ch

1 Institute of Geodesy and Photogrammetry, ETH Zurich, Robert-Gnehm-Weg 15, 8093 Zurich, Switzerland

2 Department of Geodesy, GeoForschungsZentrum (GFZ), Telegrafenberg, 14473 Potsdam, Germany

3 School of Environment Science and Spatial Informatics, China University of Mining and Technology, No. 1, Daxue Road, Xuzhou 221116, Jiangsu, China

4 Department of Geodesy and Geoinformation, TU Wien, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria

5 Navigation Science Office, European Space Agency (ESA), Camino Bajo del Castillo, 28692 Villanueva de la Cañada, Madrid, Spain

6 Novel Data Ecosystems for Sustainability Research Group, International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, 2361 Laxenburg, Lower Austria, Austria

Published online: 05 April 2024

part of the atmosphere, are split into a zenith hydrostatic delay (ZHD) and a non-hydrostatic or zenith wet delay (ZWD). The sum of ZHD and ZWD is the zenith total delay (ZTD), which amounts to roughly 2.4 m for an observer at mean sea level. The hydrostatic part makes up the majority of the ZTD and can be modelled to high accuracy with analytical methods, e.g. by using the equation of Saastamoinen (Saastamoinen 1972b). Although the wet part only contributes up to 0.40 m, it is more variable in space and time. At present there is no analytical model of ZWD with sufficient accuracy, and hence, it is typically estimated empirically.

Estimating ZWD with high accuracy is important for GNSS positioning (Ibrahim and El-Rabbany 2011; Wilgan 2015; Hadas et al. 2013), as it represents a major source of error. Furthermore, ZWD is proportional to the water vapour content along the signal path, and therefore plays an important role in GNSS meteorology (Bevis et al. 1992), with applications in weather monitoring, forecasting, and climate research (Bevis et al. 1994; Karabatić et al. 2011; Benevides et al. 2013; Seco et al. 2012; Zhao et al. 2018).

Therefore, many studies have investigated new methods to improve state-of-the-art ZWD models, such as the Hopfield model (Hopfield 1971), the Saastamoinen model (Saastamoinen 1972a), the global pressure and temperature (GPT)2w model and the GPT3 model (Böhm et al. 2015; Landskron and Böhm 2018), to name a few. Recently, machine learning (ML) approaches have also been used to construct models of tropospheric delays. Zhang et al. (2022) proposed a transformer-based global ZTD forecasting model while Yang et al. (2021) established a regional ZTD model based on the GPT3 model and artificial neural networks (ANNs). ANNs have also been used in studies by Mohammed (2021) and Selbesoglu (2020) to predict ZWDs. More recently, Ding (2022) developed a global ZWD model using neural networks that led to a better accuracy compared to state-of-the-art ZWD models (Yang et al. 2021; Böhm et al. 2015).

In this study, an ML-based model is trained based on ZWD observations of 10,718 GNSS stations during the year 2019. The reference ZWD is taken from the Nevada Geodetic Laboratory (NGL) (Blewitt et al. 2018) and the input features are the geographical location of the GNSS station, as well as the reference time epoch and meteorological variables, in particular, specific humidity on six pressure levels obtained from the ERA5 data set (Hersbach et al. 2020). The proposed model reaches centimetre-level accuracy in spatio-temporal interpolation mode, i.e. when predicting the ZWD at arbitrary spatial locations within the reference period. The temporal prediction accuracy is $\approx 2\times$ lower, but still reasonable, when extrapolating to potentially unknown atmospheric conditions outside the reference period.

Compared to the existing, relatively small-scale studies, our model is much broader. First of all, many more GNSS stations have been used to create and evaluate the established ML-based model. While Zhang et al. (2022) and Mohammed (2021) only used 505 stations, the ML-based ZWD model proposed in this study is based on 13,402 globally distributed stations (10,718 training stations and 2684 test stations). The higher number of stations leads to better performance and better generalisation of the model, especially in regions with a sparse GNSS station network. Second, in contrast to all previously mentioned ML models, our proposed ZWD model does not rely on prior ZWDs or ZWD properties to make its predictions. It is based entirely on meteorological variables, position, and time information. Therefore, the model can be applied anywhere on Earth, not only at the locations of existing GNSS stations, opening up a wide variety of applications ranging from climate research to more accurate navigation with low-cost GNSS devices, including smartphones. For the present study, the model utilised post-processed meteorological data that is available with a temporal lag of five days. Forecasts of the meteorological values are also available and could replace the reprocessed values if the model is to be used for ZWD forecasting. Here, we focus on global spatial modelling of ZWDs within a given year.

In Sect. 2, the reference ZWD as well as the meteorological variables are presented. Section 3 introduces the methodology by giving an overview of the algorithms used (Sect. 3.1), the setup (Sect. 3.2), and the validation strategy (Sect. 3.3). In Sect. 4, the ZWD predictions of the final model are shown, discussed, and thoroughly evaluated (Sects. 4.1, 4.2, 4.3, 4.4). Furthermore, several comparisons with independently computed ZWDs (Sect. 4.5) and baseline models (Sect. 4.6) have been carried out. Section 5 contains a discussion of the global model by comparing it to regional (Sect. 5.1.1) and monthly models (Sect. 5.1.2). Furthermore, the global model is applied for a different year, thus, making temporal predictions (Sect. 5.2), and the applicability of the model is explained (Sect. 5.3). In Sect. A.1 and A.2 in the appendix, the comparison of different ML algorithms and the feature selection process is further explained in more detail. Finally, Sect. 6 summarises the findings of the study and gives an outlook on future plans and further improvements.

2 Data

2.1 Zenith wet delay

Zenith wet delay (ZWD) estimates have been provided by the Nevada Geodetic Laboratory (NGL) since 1994 for a global network of GNSS stations with a temporal resolution of five minutes (Blewitt et al. 2018). NGL processes the GNSS measurements by using Jet Propulsion Laboratory's (JPL) GipsyX 1.0 software (Bertiger et al. 2020) and JPL's Repro 3.0 orbits and clocks. The tropospheric delay is calcu-

lated using the Vienna mapping function 1 (VMF1) (Boehm et al. 2006), having separate mapping functions for the ZHD and ZWD, and its gridded map products of a-priori ZHD and ZWD. Additionally, north–south and east–west gradients are estimated together with ZWD as piece-wise constants. The processing assumes a correct ZHD, with the residual delay in the ZWD. Thus, small errors in a-priori ZHD do not affect the final ZTD but might affect ZWD. The details on the data processing strategy can be found in http://geodesy.unr.edu/gps/ngl.acn.txt (last access: 25 January 2024).

For the present study, we use the ZWDs of 13,440 GNSS stations from the year 2019. To match the temporal resolution of the meteorological variables (see Sect. 2.2), the ZWD data set is down-sampled to an hourly resolution by taking the ZWD values at every full hour. This leads to a total of 117,734,400 potential samples (8760 hourly time steps $\times$ 13,440 stations). Since not all GNSS stations are recording continuously, 21,462,376 are missing, resulting in a total of 96,272,024 available samples. Although the ML algorithm is resilient against a moderate number of outliers (see Sect. 4.2), a rigorous outlier detection procedure for the ZWD data has been established. It employs four different filters: (1) The 1 % ZWD estimates with the highest uncertainties (> 3.5 mm), i.e. the standard deviations according to the product, were removed (968,173 samples). (2) Negative ZWD estimates were removed because they are physically meaningless (396,619 samples). Those estimates are likely due to ZHD modelling errors. (3) All ZWD estimates were removed that differ from the 5-hour floating median by more than $3\times$ their standard deviation (2,494,624 samples). (4) There are 573 sites that have at least two co-located stations within a distance of 1 km, covering a total of 1300 stations (roughly 10 % of all stations). For each of those sites, the median ZWD estimate per hour was calculated, and co-located stations with an offset above 5 mm from that median were removed (27 stations, 140,123 samples).

Cumulatively, the procedure flagged 3,922,694 unique outlier samples, or 4.1 % of the ZWD data set. After discarding them, we are left with 13,402 GNSS stations that still have observations (92,349,330 samples).

The distribution of the GNSS stations is illustrated in Fig. 1. It can be seen that the spatial distribution is far from homogeneous. Most stations are located in the northern hemisphere, especially in North America and Europe. However, in Asia (except for Japan, South Korea, and Nepal) the density of GNSS stations is very limited. In the southern hemisphere, the distribution of the stations is much sparser. In particular, in Africa and South America, only very few stations are available.

Figure 2 shows the completeness of the 13,402 utilised GNSS stations and the number of ZWD estimates per hour for the year 2019. The completeness is calculated by dividing the number of samples of each time series by the full number of hourly epochs in 2019 (i.e. 8760). The results show that the median is 93 %, 58 % of all stations have a completeness of over 90 % and only 17 % of all stations have a completeness of less than 50 %.

The variation of the number of ZWD estimates per hour is small throughout the year. The average number is 10,542 (out of a possible maximum of 13402) with a standard deviation of 372, which means that ZWD estimates of 79 % of all stations are available on average every hour.

Several papers have examined the quality of NGL's troposphere product by comparing it to other products. A study by Ding and Chen (2020) used the NGL troposphere data to evaluate the performance of the empirical troposphere model GPT3 and thus, assessed the accuracy of the NGL troposphere products. They compared 26 representative common stations of the International GNSS Service (IGS) and NGL and concluded that NGL's ZTD has the same accuracy as IGS's ZTD. Thus, they state that NGL's troposphere product can be used as a reference to evaluate troposphere models. In another study by Ding et al. (2022), a very high level of agreement between precipitable water vapour (PWV) data derived from radiosonde measurements and GNSS-derived PWV from NGL has been found. Since PWV can be derived directly from ZWD, it is also an indicator of the good quality of the NGL's ZWD. The characteristic differences in tropospheric delays between NGL products and numerical weather model ray-tracing are discussed in Ding et al. (2023). They found that in most regions the products correspond well, although in some high-altitude regions such as the Andes, the differences reach the cm-level. A recently published study (Yuan et al. 2023) carried out data screening of NGL's ZTD values and flagged fewer than 0.5 % of observations as outliers, which again demonstrated the good quality of NGL's troposphere product. These studies support the use of the product as a basis for a global ZWD model. Furthermore, there exists no comparable GNSS tropospheric data set in terms of dense global coverage, which is essential for our study. However, when using our ZWD predictions, the shortcomings of the NGL data must be taken into account as our model uses it as reference data.

2.2 Meteorological variables

The meteorological variables are provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) data set (Hersbach et al. 2020). ERA5 is the fifth-generation ECMWF atmospheric reanalysis of the global climate covering the period from 1940 to the present. It provides hourly estimates for a large number of atmospheric, land, and oceanic climate variables on a regular latitude, longitude grid of 0.25 degrees. The data can be accessed through

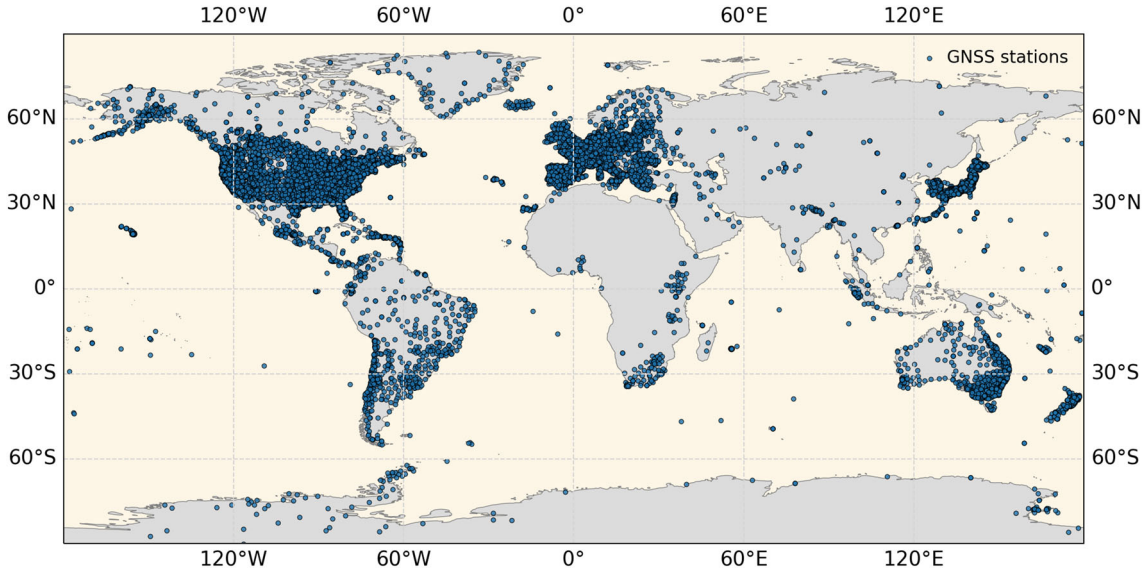

Fig. 1 Distribution of the 13,402 utilised GNSS stations

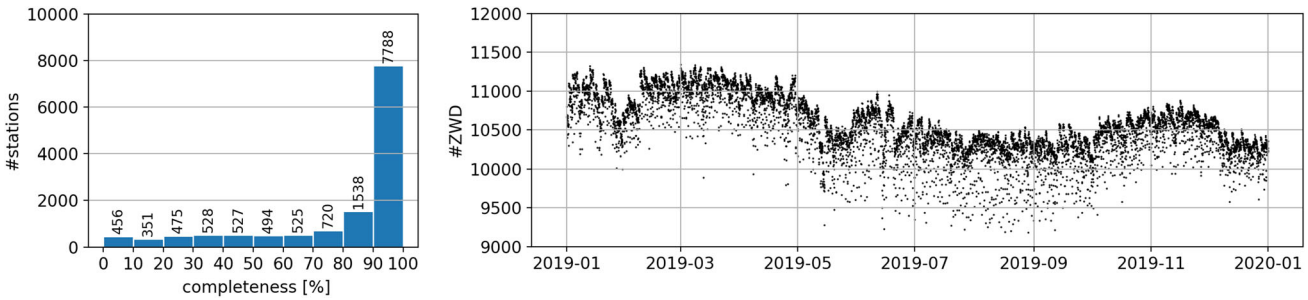

Fig. 2 Completeness of the 13,402 utilised GNSS stations (left) and the number of ZWD estimates per hour at the 13,402 utilised GNSS stations for the year 2019 (right)

the Climate Data Store[1] and are available either as single level data (roughly at surface level) or on 37 pressure levels ranging from 1000 hPa to 1 hPa. Based on expert knowledge, for our study, we selected the variables at a pressure level of 1000 hPa, except for specific humidity, where six pressure levels (1000, 900, 800, 650, 450, 300 [hPa]) are used, as further discussed in Sect. A.2 in the appendix. Table 1 lists the variables that are used within this study.

3 Methodology

In order to utilise machine learning (ML) for the determination of ZWD, three important questions have to be clarified. First, a predictive set of input features must be chosen. Second, a suitable ML algorithm has to be identified. Third, the best-performing hyper-parameters for that algorithm have to be found.

Each of these aspects depends on the others, but the huge number of possible combinations precludes an exhaustive search. We follow standard practice and take an iterative approach to determine the best-performing setup. As a start, the 13,402 stations are randomly split into 80 % training (10,718) and 20 % test stations (2684). The test stations are only utilised in the final evaluation, presented in Sect. 4, to have an independent data set. For the experiments carried out to find the best model setup, we rely only on the training stations. To that end, they are further split into five folds of equal size, four for training and one for model validation.

In the first investigation, several different ML algorithms were tested that are introduced in Sect. 3.1. In these runs, an initial set of features was selected based on expert knowledge. In total, the 12 meteorological variables at a pressure level of 1000 hPa listed in Table 1 were selected, as well as nine position and time variables describing the geographical location of the GNSS station and sample time epoch, as further discussed in Sect. 3.2.

[1] https://cds.climate.copernicus.eu/ (last access: 25 January 2024).

Table 1 Meteorological variables from the ERA5 data sets that are used in this study

Single level data	Pressure level data
Geopotential (z)	Geopotential (z_{1000})
Surface pressure (sp)	Relative humidity (r_{1000})
Total precipitation (tp)	Specific humidity ($q_{1000}, q_{900}, q_{800}, q_{650}, q_{450}, q_{300}$)
2 m temperature ($t2m$)	Temperature (t_{1000})
10 m u-component of wind (u)	U-component of wind (u_{1000})
10 m v-component of wind (v)	V-component of wind (v_{1000})

For every ML algorithm, a hyper-parameter tuning based on grid search was carried out to optimise the predictive performance of the validation set. The results of this initial comparison are presented in detail in Sect. A.1 in the appendix. Based on that investigation, the XGBoost method was found to be the most promising candidate, in line with many other ML tasks based on relatively low-dimensional feature sets (Yan et al. 2020; Lundberg et al. 2018; Hengl et al. 2017; Xia et al. 2017; Zięba et al. 2016).

In the second step, we performed a detailed feature selection for XGBoost. Several combinations of meteorological variables were studied, as further discussed in Sect. A.2 in the appendix. We found that specific humidity values at six different pressure levels, in combination with the previously mentioned representations of the position and time variables, provide good prediction performance. A full list of the features used in our final ZWD model is given in Sect. 4 in Table 2.

The hyper-parameters are then tuned again to find the best setting for the adapted feature set. It turned out that neither the hyper-parameters themselves nor the overall performance changed significantly, which demonstrates the robustness and generality of the model. The hyper-parameters that lead to the most accurate predictions are listed in Sect. 4 in Table 3.

3.1 Algorithms

To cover a broad range of ML schemes, we tested four representative methods from the vast pool of possible ML algorithms: a linear method, an exemplar-based method, a neural network, and a tree-based ensemble approach:

- Least Absolute Shrinkage and Selection Operator (LASSO) regression (Tibshirani 1996)
- K-Nearest Neighbours (KNN) (Fix and Hodges 1989; Cover and Hart 1967)
- Multilayer Perceptron (MLP) (Rosenblatt 1957; Rumelhart et al. 1986; LeCun et al. 2012)
- Extreme Gradient Boosting (XGBoost) (Chen and Guestrin 2016)

XGBoost is a tree-based ensemble learning scheme. Shallow regression trees as weak learners are combined into a strong learner with gradient boosting, i.e. the trees are sequentially learned such that they correct prediction errors of the previous stage. XGBoost is also known for its ability to capture highly nonlinear dependencies, as well as for its computational efficiency and scalability. It has achieved state-of-the-art results across a wide range of prediction tasks (Chen and Guestrin 2016).

A short description, as well as a comparison of the results of the other ML methods, can be found in Sect. A.1 in the appendix.

Additionally, two widely used methods for spatial interpolation are selected to serve as baseline models:

- Ordinary Kriging (Krige 1951)
- 3D Delaunay triangulation (Delaunay 1934)

Kriging is a spatial interpolation technique. It estimates the best linear unbiased prediction (BLUP) at an unobserved location as a weighted average of the nearby observations (Krige 1951). The weights are derived via a kernel function ("variogram") that specifies the spatial covariance structure of the target variable.

Another well-known approach to spatial interpolation in 3D is to explicitly link the 3D locations into a tetrahedral mesh with the 3D Delaunay method (Delaunay 1934), then linearly interpolate within each tetrahedron.

These baseline models were only applied to the geographical information (latitude, longitude, height) for each time step to provide a comparison with the ML models operating on the geographical as well as meteorological parameters.

3.2 Setup

As already described at the beginning of Sect. 3, all available GNSS stations are randomly split into training, validation, and test stations. For each portion, a target vector y (y_{train}, y_{val}, y_{test}) and a feature matrix X (X_{train}, X_{val}, X_{test}) are created.

Regardless of the ML method used, the learning setup in our study is always the same. The vector y, of length

Table 2 List of features utilised in the final XGBoost model

Position and time features		Meteorological features	
ϕ	Latitude	q_{1000}	Specific humidity at 1000 hPa
sin (λ)	Sine of longitude	q_{900}	Specific humidity at 900 hPa
cos (λ)	Cosine of longitude	q_{800}	Specific humidity at 800 hPa
h	Ellipsoidal height	q_{650}	Specific humidity at 650 hPa
t	Reference epoch	q_{450}	Specific humidity at 450 hPa
sin (doy)	Sine of day of year	q_{300}	Specific humidity at 300 hPa
cos (doy)	Cosine of day of year		
sin (hod)	Sine of hour of day		
cos (hod)	Cosine of hour of day		

[$\#samples$], is created by concatenating the station ZWD time series and represents the regression targets—in our case ZWD estimates from NGL. Missing data are not filled in but simply discarded.

The feature matrix X has dimension [$\#samples \times \#features$] and is composed of position and time variables (i.e. the geographical location ϕ, λ, h of the GNSS station and the sampling timestamp) and the corresponding meteorological variables, found by nearest-neighbour lookup in the ERA5 grids.

Three variables are extracted from the timestamps of each observation, which are in UTC: absolute time as a continuous, real-valued number (t); the day of the year (doy); and the hour of the day (hod). The rationale is that periodic daily and yearly signals are to be expected in ZWD time series, which are represented more directly in terms of hod and doy. To account for the cyclic nature of doy, hod, and λ, the former two are normalised to the range $[0,\ 2\pi)$ and all three are then transformed to pairs of $\sin(\cdot)$ and $\cos(\cdot)$ values, resulting in two features per variable.

Following best practice, the feature matrix X is standardised by subtracting the mean feature vector and scaling each feature dimension to unit variance, before feeding it to the ML algorithms.

3.3 Validation metrics

All quantitative results are computed on the validation fold(s) of the training set during model comparison, hyper-parameter tuning (Sect. A.1), and feature selection (Sect. A.2). Only the evaluation of the final model (Sect. 4) uses the test stations, to have an independent data set. For each test station i, we calculate the root mean squared error (RMSE; eq. (1)) and the mean absolute error (MAE; eq. (2)) between the predicted ZWD $\hat{y}$ and the (NGL-based) reference value y. As global summary statistics, the station-wise RMSEs and MAEs are combined by calculating their weighted means (WRMSE; eq. (3) and WMAE; eq. (4)), with weights proportional to the number of samples per station ($\#samples_i$). WRMSE and WMAE serve as overall performance metrics.

Table 3 Hyper-parameters of the final XGBoost model

Parameter	Value	Description
max_depth	20	Maximum tree depth for base learners
learning_rate	0.25	Boosting learning rate (shrinks the feature weights after each boosting step to make the boosting process more conservative and prevent over-fitting)
n_estimators	100	Number of gradient boosted trees

$$\mathrm{RMSE}_i = \sqrt{\frac{\sum_j^{\#samples_i} (y_{i,j} - \hat{y}_{i,j})^2}{\#samples_i}} \tag{1}$$

$$\mathrm{MAE}_i = \frac{\sum_j^{\#samples_i} \mid y_{i,j} - \hat{y}_{i,j} \mid}{\#samples_i} \tag{2}$$

$$\mathrm{WRMSE} = \frac{\sum_i^{\#stations} (\#samples_i \cdot \mathrm{RMSE}_i)}{\sum_i^{\#stations} (\#samples_i)} \tag{3}$$

$$\mathrm{WMAE} = \frac{\sum_i^{\#stations} (\#samples_i \cdot \mathrm{MAE}_i)}{\sum_i^{\#stations} (\#samples_i)} \tag{4}$$

4 Results

In this section, results for our final, best-performing global model, based on XGBoost, are presented. The features used in that model are listed in Table 2 and the hyper-parameters are given in Table 3. In the appendix, in Sects. A.1 and A.2 detailed insights are given into why XGBoost was selected and discuss how the hyper-parameters have been chosen and how the features have been selected.

4.1 Internal validation

To evaluate how well our model is able to reproduce the behaviour of the reference solution used as training target,

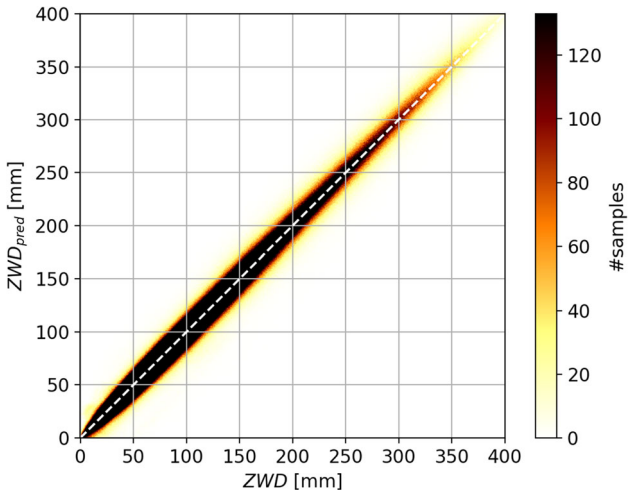

Fig. 3 Comparison of predicted ZWD values to reference values at the test stations

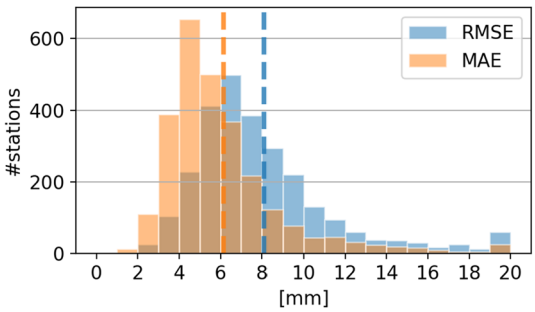

Fig. 4 Distribution of station-wise RMSE and MAE at test stations, relative to NGL reference values. Vertical lines denote WRMSE (8.1 mm) and WMAE (6.1 mm). Stations with errors larger than 20 mm are grouped in the last bin

its predictions at the test stations ($\hat{y}_{test}$) are compared to the corresponding reference values from NGL (y_{test}). In Fig. 3, XGBoost predictions of ZWD are plotted against NGL reference values. The values cluster tightly along the identity line (white, dashed) across the entire range from 0 to 400 mm ZWD, with barely any outliers. Moreover, positive and negative deviations from the ideal diagonal are symmetric, meaning that the model does not systematically over- or underpredict anywhere in the relevant range.

Figure 4 displays a histogram of the station-wise RMSE and MAE values. The distribution is skewed towards 0, meaning that most stations have small errors, while there are few stations with significantly larger errors. The weighted means of the error distributions, corresponding to WRMSE and WMAE, are 8.1 mm and 6.1 mm, respectively. Upon inspection, most of the stations with large errors (> 20 mm) are located near the coast or on islands, predominantly in tropical or subtropical regions. We speculate that in those areas the meteorological parameters may be less accurate.

The spatial distribution of the test stations' RMSE values is depicted in Fig. 5. The MAE distribution exhibits a very similar pattern and is not separately shown. A number of interesting trends can be seen. Test stations in areas with a dense GNSS station network (conterminous USA, Europe, Japan, and south-eastern Australia) tend to have lower errors. As a uniform random train/test split is used, the distribution of training stations is comparable to the one in the figure. In other words, predictive skill is better in areas with a high density of GNSS stations (and thus many training examples), as expected. This behaviour is further studied in Sect. 5.1.1, by constructing regional models. We also note that, at comparable (low) station density, the errors tend to be higher in tropical regions than in the Arctic and Antarctic, which can be explained by the much larger absolute ZWD values and their variability, another expected behaviour that we revisit in Sect. 5.1.2.

To better understand the behaviour of the ML model, the station-wise RMSE and MAE values of the test stations are related to the geographical similarity with the nearest training stations. For each test station, the Euclidean distance and the height difference to the nearest training location are determined. Then, the correlation between these values and the ZWD errors (both RMSE and MAE) are computed, see Fig. 6. The intuition behind this investigation is to see how strongly the predictive skill of the model depends on having a training station close by. All four correlation coefficients lie around $0.30-0.33$. In other words, having a nearby station does play a certain role, but the model does not just memorise the training station values (in which case the correlation with distance would have to be higher). We point out that the observed correlations are likely skewed, due to the imbalanced distribution of the distances with many more stations from areas with dense GNSS networks, and consequently also small station-to-station distances. In addition, we also computed the correlation with the absolute station height, which is very low (≤ 0.06). We speculate that several effects related to the absolute station height might cancel each other out. Overall, there are fewer stations at higher altitudes, thus fewer samples to train the model. Furthermore, maintenance of stations at higher altitudes is in general more difficult, which might affect the quality of their observations. However, at higher altitudes, the ZWDs are typically smaller and consequently have smaller errors.

4.2 Robustness against outliers

Despite the thorough outlier detection scheme described earlier (Sect. 2.1), some outliers may remain in the data set. In ML applications based on large data sets, it is not feasible to perform a manual outlier detection by individually inspecting every time series. Instead, the models are designed to be robust and/or to include automatic quality control.

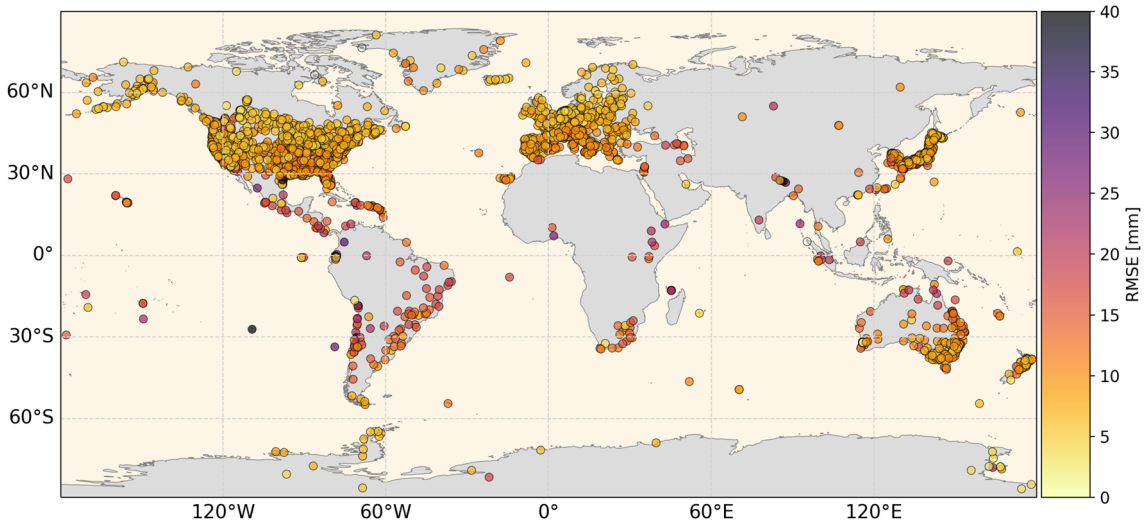

Fig. 5 Spatial distribution of the test stations' RMSEs w.r.t. NGL reference values

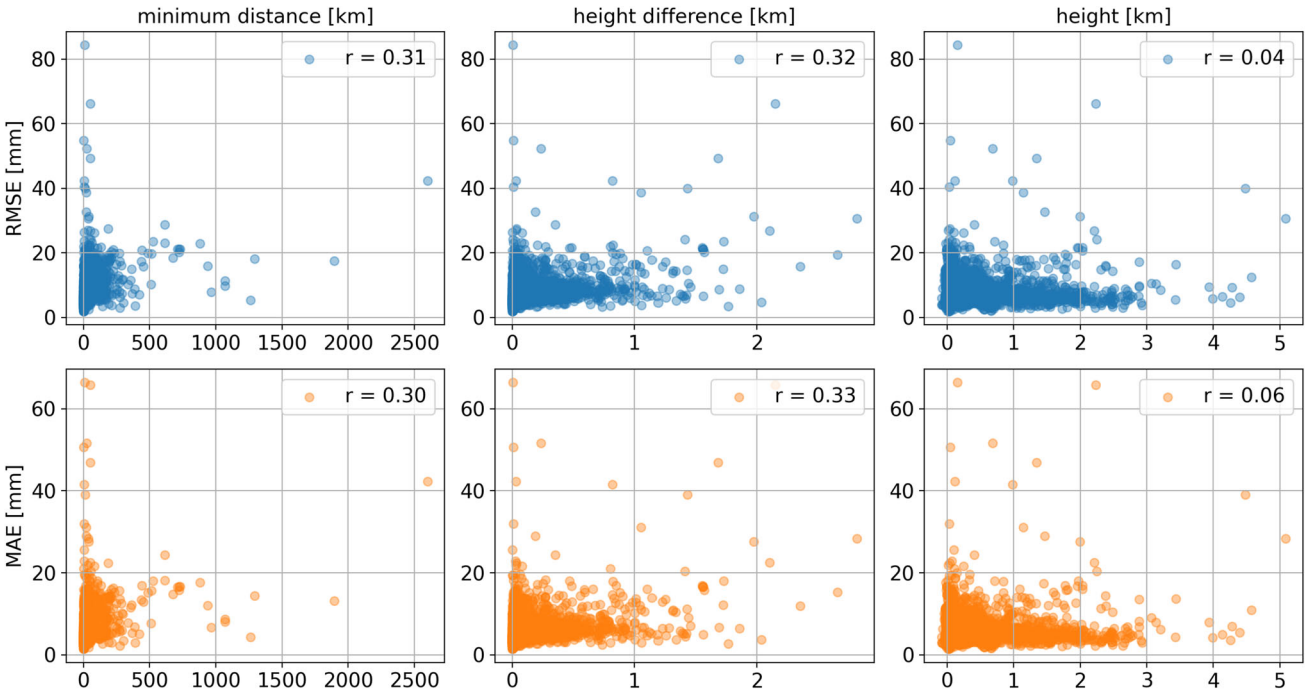

Fig. 6 RMSE and MAE at test stations, plotted against the distance to the nearest training station (left), the height difference to the nearest training station (middle), and the absolute height (right). Correlation coefficients are given in the upper right corner of each plot

To evaluate the robustness of the model against noisy data and outliers empirically, the following investigations have been conducted. First, the ZWDs of some training stations were modified to produce erroneous data. Next, a new XGBoost model was trained based on the altered ZWDs. Finally, the model was evaluated and compared to the unaltered test stations. This test was conducted for outlier rates of 1 % (107 stations) and 5 % (538 stations), significantly higher than the actual outlier ratio of NGL, which lies below 0.5 % according to previous studies (Yuan et al. 2023). The ZWDs were perturbed by various levels of white noise, and systematically biased by values between 1 and 20 mm. The artificial degradations thus reflect the entire range from small deviations that would be expected due to coordinate estimation errors up to untypically large values. Table 4 lists the WRMSE changes for all tested perturbation levels of the training data, as a way to quantify the robustness of the proposed ML approach.

For all tested white noise levels and biases, the resulting WRMSE changes are within 0.1 mm in the case of 1 % of

Table 4 Change in WRMSE based on modified ZWD training data. Positive values indicate a degradation in WRMSE. The reference WRMSE based on the unaltered ZWD is 8.1 mm

	1 mm	2 mm	5 mm	10 mm	15 mm	20 mm
White noise						
1 %	0.01 mm	0.03 mm	0.01 mm	−0.01 mm	0.04 mm	0.05 mm
5 %	−0.01 mm	0.02 mm	0.01 mm	0.05 mm	0.13 mm	0.18 mm
Bias						
1 %	−0.04 mm	0.02 mm	0.00 mm	0.02 mm	0.07 mm	0.11 mm
5 %	0.02 mm	0.01 mm	0.08 mm	0.16 mm	0.32 mm	0.55 mm

artificial outliers. Even with the exaggerated 5 % outlier tests, the changes in terms of WRMSE are within 0.2 mm for all tested white noises and biases up to 10 mm before growing to [0.3, 0.6] mm for [15, 20] mm biases, respectively. In conclusion, the ML model based on XGBoost is robust enough for the application and can deal with a reasonable amount of outliers and poor-quality data. That robustness is due to the large sample size combined with the inherent tolerance of the model to label noise in the ZWDs as well as hyper-parameters tuned on unseen data with cross-validation.

4.3 Global spatial ZWD predictions

The proposed ML model can be applied at any location on Earth and at any desired time, as long as the meteorological input variables are available. This ability is visualised by predicting global maps of ZWD with a 0.25° spatial resolution and 1-hour temporal resolution. Figure 7 shows the resulting ZWD maps for 00:00 UTC on the first day of each month in 2019.

The maps reveal the expected large-scale patterns, with overall higher values in the tropics and lower values in the polar regions. Additionally, regional weather phenomena can be distinguished, such as the South Asian monsoon that affects the Indian subcontinent from August to November. A further ZWD pattern over Central and Western Africa can probably be attributed to the seasonal displacements of the Inter-tropical Convergence Zone (ITCZ), which drives rainfall. When comparing the ZWD maps to the rainfall maps shown in the study by Dezfuli (2017), we find many similar patterns, which qualitatively corroborate the (relative) ZWD distribution predicted by our model.

4.4 Feature importance

Figure 8 illustrates the feature importance in the XGBoost model, representing the relative number of times a particular feature appears in a tree. It reveals that the three most important features are the specific humidities at pressure levels 900 hPa, 650 hPa, and 1000 hPa, highlighting that the humidity at the lower part of the atmosphere is the primary influence factor for ZWD. The most predictive pressure level of 900 hPa corresponds rather well to the 433 m average station height of the data set, suggesting that the specific humidity in the immediate environment of the station is of particular importance. Among the position and time features, ellipsoidal height (h) plays the biggest role, while latitude (ϕ) and longitude (λ) have less impact. In a dedicated experiment, the nine position and time features were omitted altogether. This roughly doubled both the RMSE and the MAE, showing that they do play a significant role, despite their relatively low feature importance (see Table 12 in Sect. A.2).

4.5 External validation

To further assess the learned model, an inter-comparison with three independent ways to estimate ZWD was performed: vertical integration of (1) ERA5 data, (2) radiosonde observations, and (3) estimation of ZWDs from the Very Long Baseline Interferometry (VLBI) analysis.

For ERA5, we obtain hourly temperature, humidity, and geopotential fields at 37 pressure levels and apply the approach of Zus et al. (2012) to the test stations. The deviations between the resulting ZWD estimates and our XGBoost model have a WRMSE of 9.1 mm and a WMAE of 6.9 mm, only 10 % higher than the values obtained in the internal validation against NGL, the reference for our study. To also quantify how well ZWDs from ERA5 and NGL agree, we compute the statistics for the difference between them and obtain a WRMSE of 10.8 mm and a WMAE of 8.3 mm for the test stations (see Table 5). Our XGBoost model thus reproduces NGL results better than a direct integration of ERA5.

In a similar fashion, ZWDs from radiosonde observations were obtained by integrating the wet refractivity vertical profiles which can be computed using pressure, temperature, and relative humidity data from radiosonde measurements (Zhang et al. 2021). The radiosonde-based ZWDs were then inter-compared to the other data sets. The radiosonde data are provided by the Integrated Global Radiosonde Archive (IGRA) (Durre et al. 2006, 2018). For the year 2019, 790 radiosonde stations are available. However, their geographical locations do not coincide with the GNSS sta-

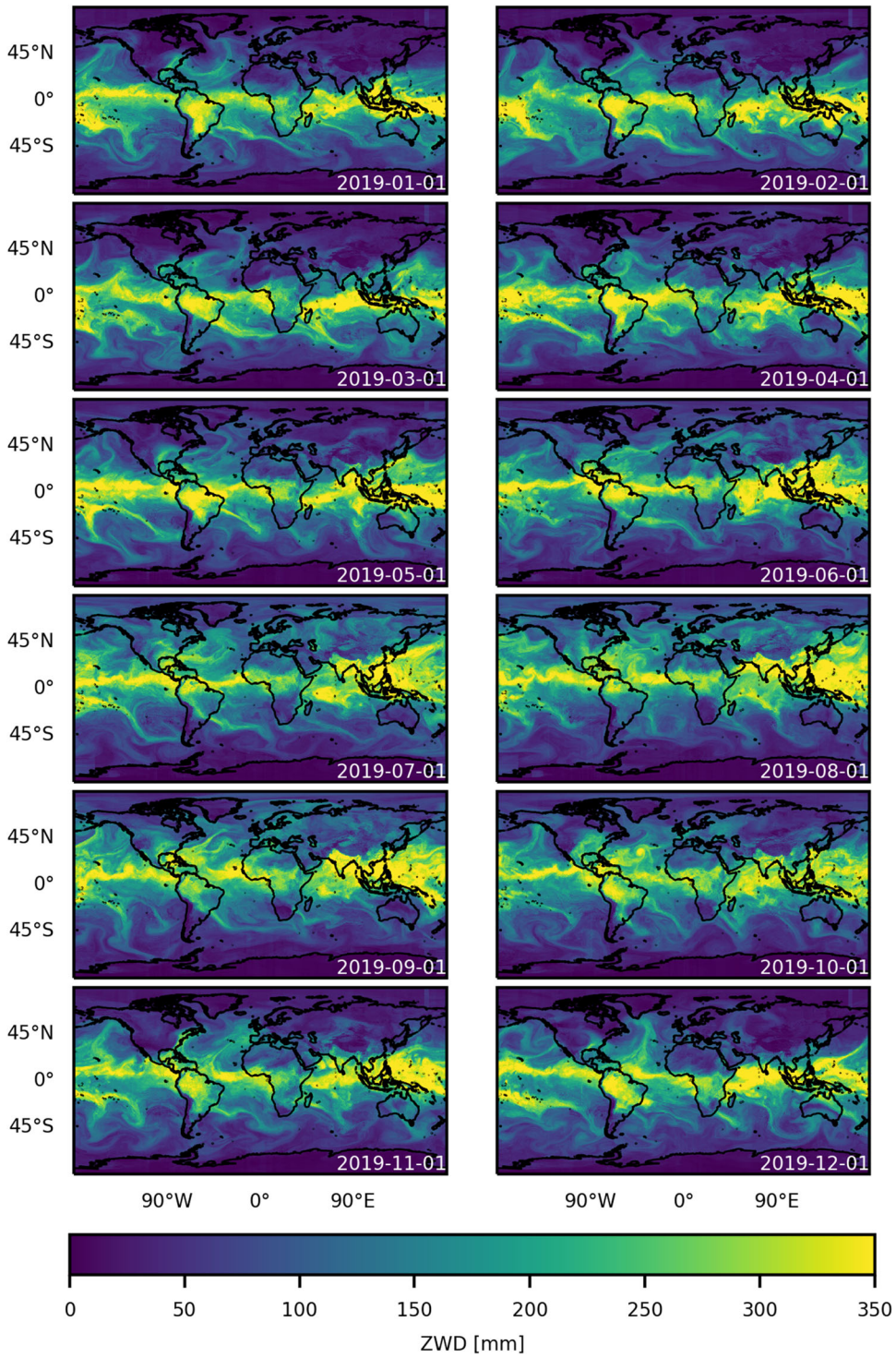

Fig. 7 Global ZWD maps for 00:00 UTC on the first day of every month in 2019

tions. To minimise the influence of spatial ZWD variability, radiosonde locations were only paired with GNSS stations if they lie within a radius of 20 km, which leaves us with 116 station pairs. At those locations, the differences between ZWDs from radiosondes and from NGL have a WRMSE of 14.5 mm and a WMAE of 11.5 mm. This result further highlights that ZWD estimates from existing retrieval methods exhibit noticeable discrepancies. The (average) deviations between the radiosonde results and our XGBoost model are very similar, with a WRMSE of 15.0 mm, respectively, a WMAE of 11.7 mm for the 116 station pairs. Finally, the deviations between radiosondes and ERA5 integration (calculated over the 116 radiosonde stations) amount to 11.0 mm WRMSE, respectively, 8.2 mm WMAE. A better agree-

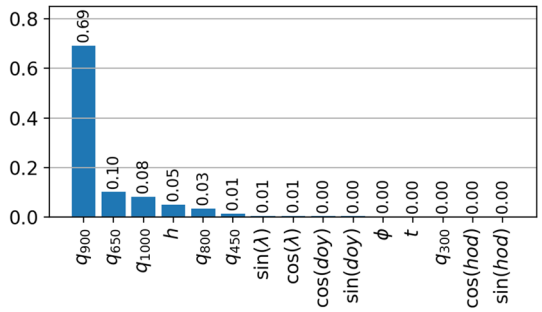

Fig. 8 Feature importance in the final ZWD model

ment of radiosonde measurements with ERA5 (than NGL or XGBoost) was expected as the radiosonde measurements are assimilated in ERA5 (Virman et al. 2021).

Additionally, ZWDs estimated during VLBI analysis are compared to ZWDs of NGL, XGBoost, and ERA5. All VLBI R1/R4 sessions of the year 2019 have been analysed based on the bkg2023a solution. In total, 104 sessions containing 24 stations (excluding the station NOTO) with 25,100 samples were included. The agreement between ZWDs of NGL and VLBI is the best with a WRMSE of 5.4 mm and a WMAE of 4.1 mm. Comparing the ZWDs of VLBI to our XGBoost model and the ZWDs of ERA5, respectively, reveals that our XGBoost model performs better (WRMSE of 10.6 mm, WMAE of 7.9 mm) than a vertical integration of ERA5 (WRMSE of 12.3 mm, WMAE of 9.0 mm). The intercomparison is summarised in Table 5. In summary, the higher compatibility of the proposed model with NGL and ERA5 was expected, given that the former served as its training target and the latter is based on the same meteorological data set. Why the discrepancies between NGL and ERA5 are lower than between either of them and the radiosonde data is less clear. This may be due to the distribution of the radiosonde locations, or it might hint at systematic observation biases.

We point out that our ML approach can, in principle, be trained with any desired ZWD data set as the regression target. Given the significant differences between retrieval methods, further research may be needed to determine to what extent training data from different sources can be mixed.

4.6 Baseline model comparisons

To compare our meteorologically informed ZWD predictions to conventional spatial interpolation techniques, we also fit ZWD fields to the set of GNSS training stations with two baseline models, Ordinary Kriging and Delaunay triangulation (see Sect. 3.1 for brief descriptions of those standard techniques).

For Ordinary Kriging, we employ the implementation available in SciKit-GStat (v.1.0.1) (Mälicke et al. 2021; Mälicke 2022). As for the previous learning algorithms, the regression targets are the ZWD values from NGL, but the input in this case is only the geographical location of the stations (latitude, longitude, height). The Delaunay interpolation is based on implementation in scipy (v.1.8.0) (Virtanen et al. 2020). Again, the training stations serve as coordinates that define the tetrahedral regions, from which the ZWD values at the test locations are read out by barycentric interpolation within the relevant tetrahedron.

The ZWD values predicted by Ordinary Kriging, respectively, Delaunay are then compared to the NGL values at the test stations, see Table 6. For the former, we obtain a WRMSE of 19.6 mm and a WMAE of 14.7 mm; for the latter, we get similar values of 18.3 mm WRMSE and 13.7 mm WMAE. These values are significantly higher than those of the XGBoost model (8.1 mm WRMSE, 6.1 mm WMAE). This was expected and confirms that the meteorological observations contribute important information about ZWD that is missing when simply interpolating the ZWD values observed at the sparse locations of the GNSS station network. The importance of meteorological data is in line with the finding of the variable importance study of Sect. A.2.

5 Discussion

The following subsections discuss the global model by comparing its performance to specialised models, namely, regional (Sect. 5.1.1) and monthly models (Sect. 5.1.2). Additionally, the global model was tested for a different year and its performance was evaluated (Sect. 5.2). Sections A.1 and A.2 in the appendix contain further details about the performance of different ML algorithms and give more insights about the feature selection.

5.1 Global versus specialised models

The final model presented in Sect. 4 is a global model that is based on 10,718 GNSS stations worldwide processed by NGL for the year 2019. In addition to creating a global model for the whole year, regional and monthly models were also generated. With these more specialised, (spatially or temporally) local models we investigate the prediction quality in more detail. Moreover, by comparing such specialised models to the monolithic, global one we are able to study the associated trade-offs. For instance, a single, global model has a larger training set and may be beneficial in regions with few stations, while on the other hand, it faces a more difficult task, as it must cover a broader range of geographical and meteorological conditions.

In the following sections, it is investigated how well the regional and monthly models performed w.r.t. the global model.

Table 5 WRMSEs [mm] (upper triangle, black) and WMAEs [mm] (lower triangle, grey) for the inter-comparison experiment. The numbers in the brackets refer to the number of stations that have been compared

	NGL	XGBoost	ERA5	VLBI	Radiosonde
NGL		8.1 (2684)	10.8 (2684)	5.4 (24)	14.5 (116)
XGBoost	6.1 (2684)		9.1 (2684)	10.6 (24)	15.0 (116)
ERA5	8.3 (2684)	6.9 (2684)		12.3 (24)	11.0 (116)
VLBI	4.1 (24)	7.9 (24)	9.0 (24)		–
Radiosonde	11.5 (116)	11.7 (116)	8.2 (116)	–	

Table 6 Comparison between ZWDs from meteorologically informed XGBoost, Delaunay interpolation, and Kriging

	NGL-based ZWD	
	WRMSE [mm]	WMAE [mm]
XGBoost-based ZWD	8.1	6.1
Delaunay-based ZWD	18.3	13.7
Kriging-based ZWD	19.6	14.7

In all cases NGL serves as the reference

5.1.1 Regional models

In total, six continental models were created that cover North America, South America, Europe, Africa, Asia, and Australia. These models were trained and evaluated with the respective subsets of the training and test sets defined for the global model. For a meaningful comparison, also the global model was evaluated separately for each continental subset of the test data. Results are shown in Table 7.

It can be seen that the best performance was achieved in Europe (WRMSE of 6.9 mm) and North America (WRMSE of 7.2 mm), the two regions with the highest number and the highest density of GNSS stations. The lowest performance was obtained for South America (WRMSE of 14.5 mm) and Africa (WRMSE of 13.4 mm), which have the lowest number of stations. We note that the results for Asia may not be representative, since both the training and the test sets are dominated by a small region comprising Japan, South Korea, and Nepal.

Overall, the performance gaps between the global model and its local counterparts are very small, indicating that generalising across the entire globe with a single model is indeed justified. In more detail, the results confirm our expectations: the local models perform slightly better in regions with enough stations, as they can fit the specific, narrower set of local conditions; but that small edge vanishes in regions with very few stations, as the global model benefits from the information contributed by the much larger set of more distant training stations. Further research is needed to comprehensively assess whether, for geographically restricted, high-accuracy applications in regions with many GNSS stations, localised models may bring a significant advantage.

5.1.2 Monthly models

Following previous studies (Sun et al. 2019; Ding 2022) that investigated variations of the model accuracy across different seasons and latitudes, we also split the training and test sets in time. As the seasonal cycles vary across the globe, we prefer not to split into somewhat arbitrary seasons, but instead, train and test a separate model for each calendar month of the year 2019. Again, the same train/test split is used as for the global model, just further subdivided into monthly subsets. The results of this experiment are summarised in Table 8.

It can be seen that a more local view tends to simplify the modelling problem: in all months the monthly models achieve slightly better performance than the global one, which is evaluated separately for each month of the test data. Moreover, the errors of the monthly models remain below the (spatially and temporally) global average error for all months except June, July, August, and September (when the largest ZWDs occur in the densely observed northern hemisphere).

Again, the differences are very small and confirm that a single, global model can capture the seasonal variations of ZWD. As before, it may be interesting to investigate in a further study how much of an advantage can be gained when the training period is extended by including the same month from multiple years.

To analyse the seasonal behaviour of ZWD predictions in different parts of the world, the performance is evaluated separately for the polar zones, the tropical zone, the northern temperate zone, and the southern temperate zone. Table 9 lists the climate zones, their latitude limits, and the number of (test) stations per zone. The results of the analysis are depicted in Fig. 9.

The errors are highest in the tropical zone but of a similar magnitude throughout the year. This makes sense since atmospheric water content and ZWDs are highest in the tropics, which on the one hand increases the magnitude of potential ZWD variations, and on the other hand, means that similar relative errors translate to higher absolute errors. Adding to that, the number of stations in the tropical zone is particularly low. It also makes sense that the errors in the tropical zone show only very little variability throughout the year, as a consequence of the stable climatic conditions without marked seasonality.

Table 7 Performance per geographical region, for both specialised regional XGBoost models and the global model

	WRMSE [mm]		WMAE [mm]		#stations	#samples
	Regional	Global	Regional	Global		
Europe	6.9	7.2	5.1	5.4	3155	21,047,736
North America	7.2	7.5	5.3	5.6	6685	45,444,795
Australia	8.8	9.2	6.6	6.9	889	6,328,393
Asia	8.7	9.1	6.7	7.1	1773	13,803,832
Africa	13.4	13.4	10.3	10.3	212	1,477,339
South America	14.5	14.7	11.4	11.6	539	3,332,638
Global average		8.1		6.1	13,402	92,349,330

Table 8 Performance per month of 2019, for both specialised monthly XGBoost models and the global model

	WRMSE [mm]		WMAE [mm]	
	Monthly	Global	Monthly	Global
January	5.9	6.3	4.7	5.0
February	6.1	6.5	4.8	5.1
March	6.3	6.7	4.9	5.2
April	6.8	7.2	5.3	5.7
May	7.6	8.0	5.9	6.3
June	8.7	9.2	6.8	7.2
July	9.4	9.9	7.3	7.7
August	9.5	10.0	7.4	7.8
September	8.6	9.1	6.7	7.1
October	7.3	7.8	5.7	6.0
November	6.4	6.8	5.0	5.4
December	6.3	6.7	4.9	5.2
All months		8.1		6.1

For the polar regions, the observed behaviour is plausible, too, with much lower errors presumably due to the dry atmosphere, and a relatively stronger seasonal signal. Due to the very low number of stations in the Arctic and Antarctic, we refrain from further interpretations.

The performance in the northern and southern temperate zones follows the expected pattern. In both zones, the errors exhibit a pronounced seasonal signal, dropping during the winter months and increasing over the summer when temperature and humidity (and thus also ZWD) are higher. The slightly higher accuracy in the northern hemisphere is likely not due to climatic influences, but explained by the much higher number of training stations.

Figure 9 raises the question if it is correct to use a random sample of test stations for the evaluation of the model. For example, if stations are selected in the north temperate zone that only observe during the summer months, significant biases might be introduced in the evaluation. However, due to the large sample size, it is unlikely that such biases appear. Furthermore, in our case, almost all stations observe year-round (see Fig. 2). Still, to ensure that no bias is present, the evaluation was additionally calculated only based on test stations with a completeness of at least 95 %. The resulting accuracy agrees at the sub-millimetre level with the one over all test stations.

5.2 Temporal predictions

As noted previously, the focus of our work lies on the global, spatially explicit modelling of ZWDs. In this context, it is important to emphasise that our current model is trained on data from the year 2019 only, and is therefore unaware of inter-annual variability. To quantify this limitation, we test its capability to predict ZWD in a different year, namely 2020. We obtain the meteorological variables for 2020 at the test stations and apply the model trained for 2019 to them. Therefore, ZWD predictions using the meteorological variables of the year 2020 are conducted.

The resulting WRMSE and WMAE values are 14.2 mm and 10.6 mm, respectively. They still lie in a very reasonable range (c.f. the inter-comparison of ZWD models in Sect. 4.5), but are nonetheless almost a factor $\times 2$ higher than those for 2019. Thus, to obtain the highest accuracy for a certain time period, it is necessary to retrain the model with the corresponding NGL data.

Figure 10 depicts the daily average RMSE over all test stations from January 2019 until December 2020. The clearly higher errors, as well as the higher variability and a sudden jump on New Year of 2020, indicate significant temporal over-fitting of the current model to the conditions of 2019. In other words, although our approach is able to model the spatial distribution of ZWD, the data from one particular year is not enough to learn a general model that covers the entire range of relevant meteorological conditions anywhere on Earth for multiple years or even decades. This is not surprising given the large inter-annual variability of the weather in large parts of the globe and the existence of major atmospheric phenomena that do not occur every year, such as the El Niño Southern Oscillation (ENSO) or the Northern Annular Mode (NAM) with their different phases and mag-

Table 9 Latitude range and number of (test) stations for the utilised climatic zones

Climate zone	Latitude	#stations	(#test stations)
Polar zones	$\phi \leq -66.5°$ or $\phi \geq 66.5°$	191	(44)
Northern temperate zone	$23.5° \leq \phi < 66.5°$	11,139	(2244)
Tropical zone	$-23.5° < \phi < 23.5°$	941	(162)
South temperate zone	$66.5 < \phi \leq 23.5°$	1131	(234)

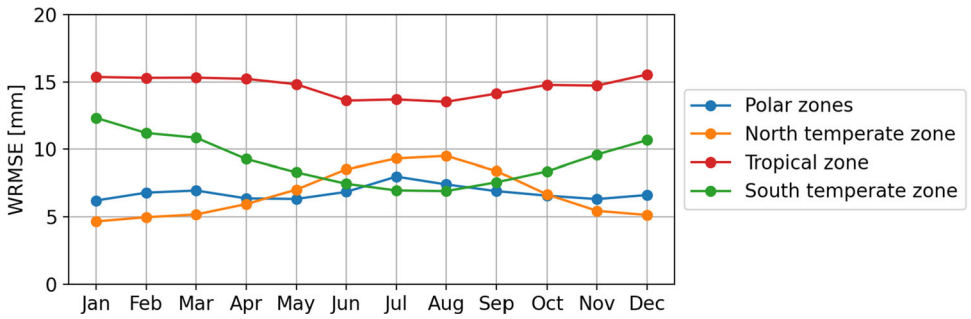

Fig. 9 Performance of monthly models evaluated in different climate zones of the world

nitudes. We point out that this shortcoming can be mitigated quite easily by extending the training data to cover multiple years (if necessary even at the cost of fewer samples per year). Additionally, the hyper-parameter tuning would have to be modified to utilise not only spatially, but also temporally independent validation data. Together, these two measures would almost certainly mitigate the problem—a promising, if obvious direction for a future extension of our model.

5.3 Applicability of our ZWD model

There are several ways in which users will be able to utilise the presented model, depending on their needs and applications.

First, a gridded data product is available that provides hourly ZWDs on a regular grid of 0.25 degrees. This data set may be useful for meteorological studies and other applications that require dense, global ZWD values. It could potentially also be used in future weather forecasting.

Second, the trained XGBoost model is provided directly. With it, users can estimate ZWD for specific locations and times but must ensure that they supply the correct inputs. Importantly, the specific humidity values at the various pressure levels are to be taken from the ERA5 data set, so as to match the data characteristics during model training. We do not recommend the use of other, user-generated specific humidity values—these would require retraining of the model. Furthermore, ZHDs have to be calculated with VMF1 to match the NGL processing of the training data to obtain realistic ZTD values.

Finally, we provide a web interface through which users can upload their location (latitude, longitude, height) and time information, which then calculates the corresponding ZWD based on the XGBoost model and ERA5 input. That interface, as well as gridded data products at 0.25° resolution are available at the Geodetic Prediction Center of ETH Zurich, https://gpc.ethz.ch/Troposphere/ (last access: 25 January 2024).

6 Conclusions and outlook

In this study, a global ML-based ZWD model is presented that achieved a performance of 8.1 mm and 6.1 mm in terms of WRMSE and WMAE, respectively, for the test stations for the year 2019. The model utilised the XGBoost algorithm with the geographical location, time epoch, and specific humidity at six pressure levels as its input features. It was trained based on hourly ZWD measurements from 10,718 GNSS stations provided by NGL for the year 2019 and evaluated against ZWD measurements from 2684 GNSS stations for the same year. The huge number of training stations ensured that the model generalised well and led to a good performance, even for regions with a sparse GNSS station network.

We verified the performance based on a thorough inter-comparison with three independent methods for determining ZWD: computation of ZWDs via (1) vertical integration of ERA5 data, (2) radiosonde measurements, and (3) ZWD estimation in VLBI analyses. Our model has the best agreement with NGL (WRMSE of 8.1 mm), which was expected since NGL serves as the reference. However, the comparison also shows good agreement between our model with ERA5 (WRMSE of 9.1 mm) and VLBI (WRMSE of 10.6 mm). The WRMSE for the radiosonde measurements is 15.0 mm. Note that the inter-comparison is based on a different number of stations (see Table 5).

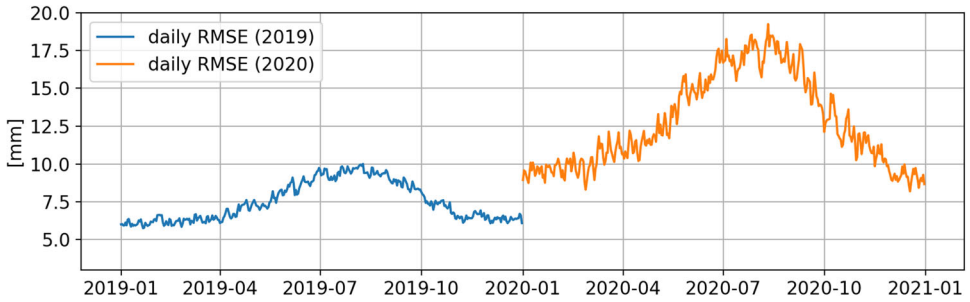

Fig. 10 Time series of daily average RMSE for 2019 (blue) and 2020 (orange)

We assume the NGL ZWD estimates to represent the ground truth. One critical question to be answered in future studies is to what extent this assumption actually holds. In particular, errors in ZHD modelling are bound to propagate into the ZWD estimates and cause local biases, which could in turn propagate into our model. A study by Ding et al. (2023) indicates that such regional biases might exist, for example, in the Andes region. Since our model has access to the geographical location, such errors would normally remain localised and not propagate to other regions. A more detailed analysis of the ZWD quality in difficult terrain is required to ascertain how the local accuracy in such regions differs from the global one. That being said, our study nevertheless demonstrates that the proposed model delivers ZWD values globally and with high accuracy in most regions of the Earth. We also note that updating the model is straightforward: if different, better ZWD values for sufficiently many reference stations become available, all one has to do is retrain the model with those values.

To further demonstrate the quality of the global model, regional (continental) and monthly models were also investigated, which showed that the differences between the WRMSE and WMAE were very small, on average 0.3 mm for the regional and 0.4 mm for the monthly models. This indicates that the global model performs reasonably well for all regions of the Earth and over the full year. Concerning the regional models, it is shown that areas with a dense GNSS station network and a high number of stations (e.g. Europe, North America) have a better performance than areas with a sparse network and a low number of stations (e.g. South America, Africa). Concerning the monthly models, it is revealed that the ZWD accuracy of stations located in the northern and southern temperate zones is worse during the corresponding summer months, likely explained by the higher water vapour content and thus higher variations in ZWDs.

One major advantage of the proposed model is that, in contrast to other ZWD models, it does not depend on prior ZWDs. Thus, it can be applied anywhere on Earth, opening up the possibility to use it for a wide range of applications in the field of positioning and possibly also for weather monitoring and forecasting. Furthermore, once trained, calculating ZWDs based on the input features is computationally inexpensive making it attractive for low-cost or low-power devices. These properties, together with the better performance compared to ZWD computed from ERA5, make the ML model superior to alternative options.

While our model was designed for spatial modelling, additional experiments were conducted regarding its potential for temporal predictions. We found that the performance noticeably drops when applying the model to data outside the training period. This can be explained by the fact that it is only trained on data from one year, and therefore unaware of inter-annual variability. To overcome this limitation, it will be necessary to train the model on multiple years in a future extension of our model, including the choice of a temporally independent validation set. We are confident that in this way temporal generalisation can be achieved, leading to improved predictions at previously unobserved points in time. In addition, this study used specific humidity from ERA5 reanalysis data that only become available within five days of the present day, which allows to present the concept of our model but is a limitation for real-time applications. Preliminary investigations towards a temporal forecasting model of ZWD suitable for real-time applications have already been presented in Crocetti et al. (2023) at the European Geoscience Union (EGU) 2023. It appears that feeding an adapted ML model with meteorological forecast data from the ECMWF Integrated Forecasting System (IFS), in combination with training data from multiple years improves ZWD estimation across time and does not lead to a significant performance loss. However, this real-time setting is still under investigation and will be reported in future works.

Acknowledgements We would like to express our gratitude to the reviewers and editors who provided valuable comments that helped us to further improve our paper. In addition, we would like to thank all the CAMALIOT team members for the many fruitful discussions and the data providers who are essential to carry out the investigations presented in this paper.

Author Contributions LC and BS designed the study with the help of MS, GM, and VN. MS performed the data pre-processing and cleaning. LC implemented the algorithms, performed the analysis, prepared the visualisations, and wrote the majority of the manuscript. The intermediate results were continuously discussed with MS and regularly presented to GM, VN, LS, KS, and BS. The radiosonde-based ZWDs

were computed by WZ. The ERA5-based ZWDs were computed by FZ. Comparison with VLBI ZWDs was performed by MS. BS and KS supervised the work. All authors discussed the results, contributed to the final manuscript, and agreed to the published version.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich. This work has been carried out as part of the project "CAMALIOT: Application of Machine Learning Technology for GNSS IoT Data Fusion", funded by the European Space Agency (ESA) NAVISP Element 1 (NAVISP-EL1−038.2).

Data availability The zenith wet delays were downloaded from the Nevada Geodetic Laboratory (NGL) (Blewitt et al., 2018).
The ERA5 data set (Hersbach et al., 2018) was downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store. The ERA5 data set was downloaded using Copernicus Climate Change Service information 2022. Neither the European Commission nor the ECMWF is responsible for any use that may be made of the Copernicus information or the data it contains. The XGBoost-based ZWD predictions are available at the Geodetic Prediction Center of ETH Zurich (https://gpc.ethz.ch). The radiosonde data was downloaded from the Integrated Global Radiosonde Archive (IGRA) (Durre et al., 2006, 2018). The VLBI data was downloaded from the Crustal Dynamics Data Information System (CDDIS).

Declarations

Conflict of interest The authors declare no conflict of interest.

Appendix A

A.1 Comparison of different algorithms

In order to identify the most suitable ML algorithm, different ML algorithms (LASSO regression, KNN, and MLP), listed in Sect. 3.1, were tested based on the validation stations.

LASSO regression is a linear regressor with L1 regularisation of the coefficients. The L1 norm encourages sparsity, i.e. coefficients with low influence on the prediction shrink to zero, thus offering automatic feature selection. A benefit of LASSO is that, due to the simplicity of the method, the resulting model is interpretable. Its most serious limitation is that only linear relationships can be modelled.

The KNN algorithm is a nonparametric method, which finds a predefined number of neighbours in feature space and computes the prediction by local interpolation of the targets. KNN belongs to the family of "lazy learners": instead of abstracting the training data into a discriminative function, it stores them directly and scans the entire training set at inference time. A downside of KNN is that, while no training is necessary at all, it needs a lot of memory, especially when processing big data sets. Additionally, KNN tends to degrade as more predictors are used because distances between points become increasingly similar and are no longer discriminative in high-dimensional spaces (Beyer et al. 1999). Dimensionality reduction techniques such as principal component analysis (Pearson 1901) can mitigate this problem, but only if the data form a low-dimensional manifold in feature space. One benefit of KNN compared to other algorithms such as LASSO is that KNN is able to solve nonlinear problems.

The MLP is a feed-forward neural network with one or more hidden layers between the input features and the output (target) value. The possibility to employ different (element-wise) activation functions between layers, and to vary the number and width of hidden layers, makes the MLP a very flexible and powerful tool, which however requires careful tuning, as well as some experience to ensure the gradient-based, highly stochastic optimisation converges properly. They are a popular choice to learn highly nonlinear relationships from large data sets.

The performance of the individual algorithms in terms of WRMSE and WMAE is depicted in Table 10.

Table 10 Performance of different ML algorithms on the validation stations

	WRMSE [mm]	WMAE [mm]
LASSO	31.2	24.1
KNN	14.7	10.5
MLP	20.2	15.4
XGBoost	12.2	9.1

The features used to model ZWD are selected based on expert knowledge (12 meteorological variables at a pressure level of 1000 hPa and nine position and time variables listed in Table 1) (see Sect. A.2) and the set of hyper-parameters for the individual algorithms are shown in Table 11. The python packages Scikit-learn (v1.0.2) (Pedregosa et al. 2011) and XGBoost (v1.5.2 and v1.6.2) (Chen and Guestrin 2016) were used for the computations. The models were trained based on 80 % of the training stations (four folds) and evaluated based on the remaining 20 % of the training stations (one fold), i.e. the validation set.

It can be clearly seen that a linear model, like LASSO, performed the worst with a WRMSE of 31.7 mm. The KNN algorithm achieved a significantly better result with a WRMSE of 14.7 mm. A more complex model, like the MLP,

Table 11 Tuned hyper-parameters for the LASSO, KNN, and MLP algorithm

ML algorithm	Parameter	Value	Description
LASSO	fit_intercept	True	Whether to calculate the intercept for this model
	max_iter	1000	The maximum number of iterations
	cv	None	The default fivefold cross-validation is used
KNN	n_neighbours	5	Number of neighbours to use
	weights	'distance'	Weight function used in prediction
	algorithm	'kd_tree'	Algorithm used to compute the nearest neighbours
MLP	hidden_layer_sizes	(256,128,64,32,16,8,4)	The ith element represents the number of neurons in the ith hidden layer
	early_stopping	True	Whether to use early stopping to terminate training when validation score is not improving
	validation_fraction	0.2	The proportion of training data to set aside as validation set for early stopping
	n_iter_no_change	30	Maximum number of epochs to not meet tolerance improvement

Table 12 XGBoost results of the validation stations of the different feature constellations

	WRMSE [mm]	WMAE [mm]
baseline features	22.3	15.7
baseline features + *meteorological variables*	12.2	9.1
baseline features + q_{1000}	12.5	9.3
baseline features + q_{six}	8.5	6.4
baseline features + q_{six} + *meteorological variables*	8.3	6.3
q_{six}	16.1	12.4

achieved a WRMSE of 20.2 mm. It is important to note that although a sophisticated hyper-parameter tuning for MLP was performed, it is likely that some more fine-tuning of the MLP model could result in a slightly better performance. However, the best-performing model found within this investigation was XGBoost with a WRMSE of 12.6 mm. Based on these results, XGBoost was selected and further investigated.

A.2 Feature selection

For the first analyses, position and time features (i.e. ϕ, λ, h of the GNSS station and time information t, doy, hod), further denoted as *baseline features*, and 12 meteorological variables (all at a pressure level of 1000 hPa), further denoted as *meteorological variables*, were selected as features based on expert knowledge. This feature set led to a WRMSE of 12.2 mm for XGBoost, as presented in Sect. A.1. In the second step, after the ML algorithm was fixed, a detailed feature selection was carried out with XGBoost. The results of the validation stations are presented in Table 12.

When using only the *baseline features*, the model performance decreased significantly to 22.3 mm of WRMSE. This indicates that the *meteorological variables* are important and need to be included.

To find out which features are the most important, a feature importance plot was created for the model that uses the *baseline features* and the *meteorological variables* and is shown in Fig. 11. This plot reveals that specific humidity (q_{1000}) is by far the most important feature, followed by total precipitation (tp) and the position and time features.

Based on these findings, another model run was carried out, to investigate whether using only the *baseline features* and specific humidity (q_{1000}) can achieve the same performance as using all *meteorological variables*. This would reduce the amount of data and speed up the model training. It turned out that this feature set reaches almost the accuracy of the model that included all *meteorological variables* (WRMSE of 12.5 mm).

An additional experiment was then run in which different pressure levels for specific humidity were included. Six

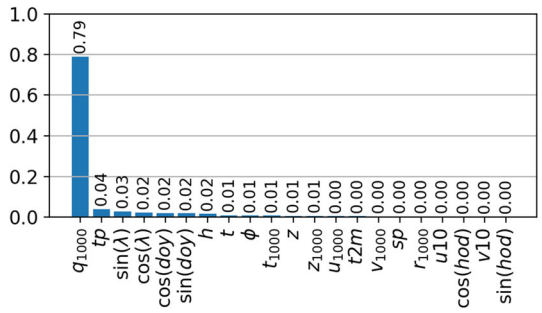

Fig. 11 Feature importance in the XGBoost model trained with *baseline features* and *meteorological variables* as inputs

Table 13 Results of the fivefold cross-validation using XGBoost and the final feature constellation (*baseline features* + q_{six})

validation	WRMSE [mm]	WMAE [mm]
fold 1	8.47	6.41
fold 2	8.48	6.38
fold 3	8.39	6.31
fold 4	8.51	6.41
fold 5	8.49	6.41

pressure levels reaching from 1000 hPa to 300 hPa (q_{six}) were selected since the water vapour is mostly located in this part of the atmosphere. The model run using this feature constellation produced significantly better results achieving a WRMSE of 8.5 mm.

Additionally, another model run was carried out, using the *baseline features*, all six pressure levels for specific humidity (q_{six}), and the remaining *meteorological variables*. The WRMSE of this run is 8.3 mm, which is the best result. However, the improvement is insignificant (only 2 %) and 11 more features were used, resulting in a much larger feature matrix that is slower to process.

As a last check, a model was created using only specific humidity on six pressure levels (q_{six}) as features and discarding the *baseline features*. The WRMSE increased significantly to 16.1 mm. Therefore, the final features are the nine *baseline features*, as well as specific humidity on six pressure levels (q_{six}).

To assess the stability of the XGBoost model with the final feature set, we ran a fivefold cross-validation. The training data was randomly split into five equally sized folds and each fold in turn was used as a validation set while training on the four remaining ones. Table 13 lists the individual performance per fold. The WRMSE and WMAE are practically the same in all five runs, showing that the method is unaffected by the exact choice of training stations and that its prediction performance is stable across varying validation sets.

References

Benevides P, Catalão J, Miranda P, et al (2013) Analysis of the relation between GPS tropospheric delay and intense precipitation. In: Comeron A, Kassianov EI, Schäfer K, et al (eds) Remote Sensing of Clouds and the Atmosphere XVIII; and Optics in Atmospheric Propagation and Adaptive Systems XVI, vol 8890. SPIE, Dresden, Germany, p 88900Y, https://doi.org/10.1117/12.2028732

Bertiger W, Bar-Sever Y, Dorsey A et al (2020) GipsyX/RTGx, a new tool set for space geodetic operations and research. Adv Space Res 66(3):469–489. https://doi.org/10.1016/j.asr.2020.04.015

Bevis M, Businger S, Herring TA et al (1992) GPS meteorology: remote sensing of atmospheric water vapor using the global positioning system. J Geophys Res Atmos 97(D14):15,787-15,801. https://doi.org/10.1029/92JD01517

Bevis M, Businger S, Chiswell S et al (1994) GPS Meteorology: Mapping Zenith Wet Delays onto Precipitable Water. J Appl Meteorol Climatol 33(3):379–386

Beyer K, Goldstein J, Ramakrishnan R et al (1999) When Is "Nearest Neighbor" Meaningful? In: Beeri C, Buneman P (eds) Database Theory – ICDT'99. Springer, Berlin, pp 217–235

Blewitt G, Hammond W, Kreemer C (2018) Harnessing the GPS data explosion for interdisciplinary science. Eos. https://doi.org/10.1029/2018eo104623

Boehm J, Werl B, Schuh H (2006) Troposphere mapping functions for GPS and very long baseline interferometry from European Centre for Medium-Range Weather Forecasts operational analysis data. J Geophys Res Solid Earth. https://doi.org/10.1029/2005jb003629

Böhm J, Möller G, Schindelegger M et al (2015) Development of an improved empirical model for slant delays in the troposphere (GPT2w). GPS Solut 19(3):433–441. https://doi.org/10.1007/s10291-014-0403-7

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. Association for Computing Machinery, New York, NY, USA, KDD '16, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27. https://doi.org/10.1109/TIT.1967.1053964

Crocetti L, Schartner M, Schindler K et al (2023). Forecasting of tropospheric parameters using meteorological data and machine learning. https://doi.org/10.5194/egusphere-egu23-3453

Delaunay B et al (1934) Sur la sphere vide. Izv Akad Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk 7(793–800):1–2

Dezfuli A (2017) Climate of Western and Central Equatorial Africa. https://doi.org/10.1093/acrefore/9780190228620.013.511

Ding J, Chen J (2020) Assessment of empirical troposphere model GPT3 based on NGL's global troposphere products. Sensors. https://doi.org/10.3390/s20133631

Ding J, Chen J, Tang W et al (2022) Spatial and temporal variability of global GNSS-derived precipitable water vapor (1994–2020) and climate implications. Remote Sens. https://doi.org/10.3390/rs14143493

Ding J, Chen J, Wang J et al (2023) Characteristic differences in tropospheric delay between Nevada Geodetic Laboratory products and NWM ray-tracing. GPS Solut 27(1):47. https://doi.org/10.1007/s10291-022-01385-2

Ding M (2022) Developing a new combined model of zenith wet delay by using neural network. Adv Space Res 70(2):350–359. https://doi.org/10.1016/j.asr.2022.04.043

Durre I, Vose RS, Wuertz DB (2006) Overview of the integrated global radiosonde archive. J Clim 19(1):53–68. https://doi.org/10.1175/JCLI3594.1

Durre I, Yin X, Vose RS et al (2018) Enhancing the data coverage in the integrated global radiosonde archive. J Atmos Oceanic Tech 35(9):1753–1770. https://doi.org/10.1175/JTECH-D-17-0223.1

Fix E, Hodges JL (1989) Discriminatory analysis. Nonparametric discrimination: consistency properties. Int Stat Rev 57(3):238–247

Hadas T, Kaplon J, Bosy J et al (2013) Near-real-time regional troposphere models for the GNSS precise point positioning technique. Meas Sci Technol 24(5):055,003. https://doi.org/10.1088/0957-0233/24/5/055003

Hengl T, de Jesus JM, Heuvelink GBM et al (2017) SoilGrids250m: global gridded soil information based on machine learning. PLoS ONE 12(2):e0169,748. https://doi.org/10.1371/journal.pone.0169748

Hersbach H, Bell B, Berrisford P et al (2020) The ERA5 global reanalysis. Q J R Meteorol Soc 146(730):1999–2049. https://doi.org/10.1002/qj.3803

Hopfield H (1971) Tropospheric effect on electromagnetically measured range: prediction from surface weather data. Radio Sci 6(3):357–367. https://doi.org/10.1029/RS006i003p00357

Ibrahim HE, El-Rabbany A (2011) Performance analysis of NOAA tropospheric signal delay model. Meas Sci Technol 22(11):115,107. https://doi.org/10.1088/0957-0233/22/11/115107

Karabatić A, Weber R, Haiden T (2011) Near real-time estimation of tropospheric water vapour content from ground based GNSS data and its potential contribution to weather now-casting in Austria. Adv Space Res 47(10):1691–1703. https://doi.org/10.1016/j.asr.2010.10.028

Krige DG (1951) A statistical approach to some mine valuation and allied problems on the Witwatersrand. University of Witwatersrand, Johannesburg

Landskron D, Böhm J (2018) VMF3/GPT3: refined discrete and empirical troposphere mapping functions. J Geodesy 92(4):349–360. https://doi.org/10.1007/s00190-017-1066-2

LeCun YA, Bottou L, Orr GB, et al (2012) Efficient backProp. Springer, Berlin, pp 9–48. https://doi.org/10.1007/978-3-642-35289-8_3

Lundberg SM, Nair B, Vavilala MS et al (2018) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2(10):749–760. https://doi.org/10.1038/s41551-018-0304-0

Mälicke M (2022) SciKit-GStat 1.0: a SciPy-flavored geostatistical variogram estimation toolbox written in Python. Geosci Model Dev 15(6):2505–2532. https://doi.org/10.5194/gmd-15-2505-2022

Mälicke M, Möller E, Schneider HD, et al (2021) mmaelicke/scikit-gstat: a scipy flavoured geostatistical variogram analysis toolbox. https://doi.org/10.5281/zenodo.4835779

Mohammed J (2021) Artificial neural network for predicting global sub-daily tropospheric wet delay. J Atmos Solar Terr Phys 217(105):612. https://doi.org/10.1016/j.jastp.2021.105612

Nilsson T, Böhm J, Wijaya DD, et al (2013) Path delays in the neutral atmosphere. Springer, Berlin, pp 73–136. https://doi.org/10.1007/978-3-642-36932-2_3

Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. Lond Edinb Dublin Philos Mag J Sci 2(11):559–572. https://doi.org/10.1080/14786440109462720

Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Rosenblatt F (1957) The perceptron, a perceiving and recognizing automaton (Project Para). Cornell Aeronautical Laboratory, Buffalo, NY

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536. https://doi.org/10.1038/323533a0

Saastamoinen J (1972a) Atmospheric correction for the troposphere and stratosphere in radio ranging satellites. In: The use of artificial satellites for geodesy. American Geophysical Union (AGU), pp 247–251. https://doi.org/10.1029/GM015p0247

Saastamoinen J (1972b) Contributions to the theory of atmospheric refraction. Bull Geodesique (1946–1975) 105(1):279–298. https://doi.org/10.1007/BF02521844

Seco A, Ramírez F, Serna E et al (2012) Rain pattern analysis and forecast model based on GPS estimated atmospheric water vapor content. Atmos Environ 49:85–93. https://doi.org/10.1016/j.atmosenv.2011.12.019

Selbesoglu MO (2020) Prediction of tropospheric wet delay by an artificial neural network model based on meteorological and GNSS data. Eng Sci Techno Int J 23(5):967–972. https://doi.org/10.1016/j.jestch.2019.11.006

Sun Z, Zhang B, Yao Y (2019) A global model for estimating tropospheric delay and weighted mean temperature developed with atmospheric reanalysis data from 1979 to 2017. Remote Sens. https://doi.org/10.3390/rs11161893

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B (Methodol) 58(1):267–288

Virman M, Bister M, Räisänen J et al (2021) Radiosonde comparison of ERA5 and ERA-Interim reanalysis datasets over tropical oceans. Tellus A Dyn Meteorol Oceanogr 73(1):1–7. https://doi.org/10.1080/16000870.2021.1929752

Virtanen P, Gommers R, Oliphant TE et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2

Wilgan K (2015) Zenith total delay short-term statistical forecasts for GNSS precise point positioning. Acta Geodynamica et Geomaterialia. https://doi.org/10.13168/agg.2015.0035

Xia Y, Liu C, Li Y et al (2017) A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Syst Appl 78:225–241. https://doi.org/10.1016/j.eswa.2017.02.017

Yan L, Zhang HT, Goncalves J et al (2020) An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell 2(5):283–288. https://doi.org/10.1038/s42256-020-0180-7

Yang F, Guo J, Zhang C et al (2021) A regional zenith tropospheric delay (ZTD) model based on GPT3 and ANN. Remote Sens. https://doi.org/10.3390/rs13050838

Yuan P, Blewitt G, Kreemer C et al (2023) An enhanced integrated water vapour dataset from more than 10 000 global ground-based GPS stations in 2020. Earth Syst Sci Data 15(2):723–743. https://doi.org/10.5194/essd-15-723-2023

Zhang H, Yao Y, Xu C et al (2022) Transformer-based global Zenith tropospheric delay forecasting model. Remote Sens. https://doi.org/10.3390/rs14143335

Zhang W, Zhang S, Zheng N et al (2021) A new integrated method of GNSS and MODIS measurements for tropospheric water vapor tomography. GPS Solut 25(2):79. https://doi.org/10.1007/s10291-021-01114-1

Zhao Q, Yao Y, Yao W et al (2018) Real-time precise point positioning-based zenith tropospheric delay for precipitation forecasting. Sci Rep 8(1):7939. https://doi.org/10.1038/s41598-018-26299-3

Zięba M, Tomczak SK, Tomczak JM (2016) Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Syst Appl 58:93–101. https://doi.org/10.1016/j.eswa.2016.04.001

Zus F, Bender M, Deng Z et al (2012) A methodology to compute GPS slant total delays in a numerical weather model. Radio Sci. https://doi.org/10.1029/2011RS004853