GPS Solutions (2025) 29:59
https://doi.org/10.1007/s10291-024-01796-3

ORIGINAL ARTICLE

An advanced regional integrated water vapor estimation model utilizing least squares support vector machine for the upper Rhine graben region

Lingke Wang[1] · **Hansjörg Kutterer**[1]

Received: 20 June 2024 / Accepted: 29 November 2024 / Published online: 30 January 2025

Abstract
Integrated Water Vapor (IWV) is crucial in environmental research, offering insights into atmospheric dynamics. Direct IWV measurement is challenging, necessitating alternative estimation technologies. Existing methods including Global Navigation Satellite System (GNSS), radiosondes, water vapor radiometers (WVR), satellite remote sensing, and numerical weather models (NWM), have specific limitations. GNSS and WVR provide high precision and temporal resolution (e.g., 5 min) but are limited to specific locations. Radiosondes, while accurate, have sparse spatial distribution and low temporal resolution (e.g., twice daily). Satellite remote sensing offers broad spatial resolution but lower temporal resolution (hours to days) and reduced accuracy under cloudy conditions and due to satellite tracks. NWMs provide global hourly products but their accuracy depends on meteorological data and model precision.
This study introduces a regional IWV predictive model using Machine Learning to address these challenges. Utilizing IWV data from GNSS stations, the study develops a predictive model based on least squares support vector machine, which autonomously determines optimal parameters to enhance performance. The model enables accurate IWV estimation at any location within a region, using inputs such as latitude, longitude, altitude, and temperature, achieving an average root mean square error of 0.95 mm. The model's performance varies across seasons and terrains, showing adaptability to diverse conditions. The model's reliability is validated by comparing its predictions with the conventional ERA5 IWV method, showing a 61% improvement rate. This refined IWV estimation model is applied for regional climate analysis, demonstrating its practical utility in environmental research, specifically for the Upper Rhine Graben Region.

Keywords Global navigation satellite system (GNSS) · Integrated water vapor (IWV) · Least squares support vector machine (LSSVM) · IWV predictive model

Introduction

The troposphere encompasses the majority of the atmospheric mass and water vapor content. Water vapor is predominantly found in the troposphere, typically below 10–12 km altitude. This concentration is pivotal for meteorological investigations, encompassing studies on atmospheric dynamics, weather prediction, and climate modeling (Yuan et al. 2022). Moreover, precise Integrated Water Vapor (IWV) datasets are indispensable for remote sensing applications. Enhanced IWV accuracy improves the dependability and precision of remote sensing outputs, spanning satellite imagery, weather radar data, and atmospheric profiling instruments (Izanlou et al. 2024).

Traditionally, IWV measurements have been conducted through various in-situ techniques such as Radiosonde and Radiometer. While these methods exhibit high accuracy, their spatial coverage is constrained, being available only at specific stations rather than across expansive regions. Furthermore, their installation and establishment necessitate considerable human and financial resources (Seidel et al. 2011; Migliaccio et al., 2005). In contrast, the European Centre for Medium-Range Weather Forecasts (ECMWF) offers global grid numerical weather models (NWM), such as the fifth generation of European Reanalysis (ERA5), which can be used to estimate large-scale IWV based on the

✉ Lingke Wang
lingke.wang@kit.edu

[1] Geodätisches Institut, Karlsruher Institut Für Technologie, Englerstraße 7, 76131 Karlsruhe, Germany

interpolation of pressure-level meteorological parameters and zenith wet delay (ZWD) ray tracing models (Ssenyunzi et al. 2020; Wilgan et al. 2017). These models then convert ZWD to IWV by integrating the weighted mean temperature (Bevis et al. 1994). However, despite providing large-scale IWV products, their performance is limited by the accuracy of the ZWD ray tracing model and the challenge of accurately modeling IWV variations through interpolated meteorological elements (Huang et al. 2022). Therefore, the primary focus of our research is to develop an enhanced IWV estimation model capable of operating at a large scale while still delivering highly accurate IWV products. This entails addressing the limitations of existing methodologies and improving upon them to achieve greater precision and broader applicability across expansive geographical areas.

The rapid evolution of Global Navigation Satellite System (GNSS) technology has positioned it as a crucial instrument for tracking water vapor variations (Rocken et al. 1993). As GNSS signals propagate through the troposphere, they encounter distortions and deflections caused by atmospheric refraction. This phenomenon is commonly assessed as the zenith tropospheric delay (ZTD), wherein the calculation involves the mapping function upon the elevation angle utilized during GNSS data analysis (Askne et al., 1987; Böhm et al. 2006a, b). The tropospheric delay comprises two components: the zenith hydrostatic delay (ZHD) and ZWD. Water vapor variations can be derived from the ZWD, which is the difference between ZTD and ZHD. ZHD can be accurately computed using various estimation models, such as the Saastamoinen model (Saastamoinen 1972). Thus, with high-precision ZHD data, accurate ZWD can be derived, and by incorporating temperature data, ZWD can be converted into IWV, allowing for precise IWV monitoring (Jiang et al. 2020). GNSS technology offers the advantage of retrieving IWV data with higher accuracy, supported by its dense distribution of sites worldwide. Furthermore, GNSS derived IWV data exhibit high temporal resolution (Baba et al., 2022). Given these capabilities, our research pursuits to exploit isolated GNSS IWV time series from different stations to extrapolate IWV values across a broader geographical area. Consequently, this study aims to develop an enhanced regional IWV estimation model based on simultaneous GNSS IWV values, enabling the estimation of IWV at any location and altitude within this large-scale region.

This study opts for data sourced from the Nevada Geodetic Laboratory (NGL) for testing within the Upper Rhine Graben Region. NGL has made substantial strides by consolidating global GNSS observations from over 10,000 stations and offering ZTD estimates (Blewitt et al., 2018). NGL uniformly processes GNSS observations using contemporary models and methodologies. Leveraging ZTD data from NGL, our research aims to convert ZTD into IWV and establish a regional GNSS network comprising approximately 150 stations, which are represented in Fig. 1.

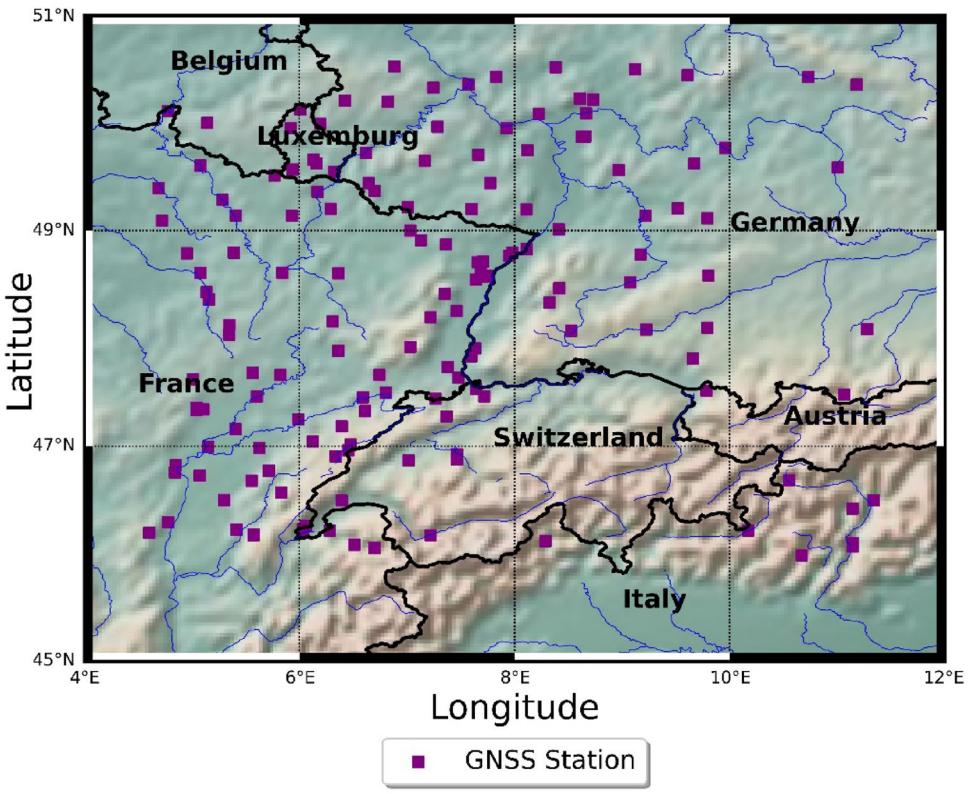

Fig. 1 GNSS stations distribution map

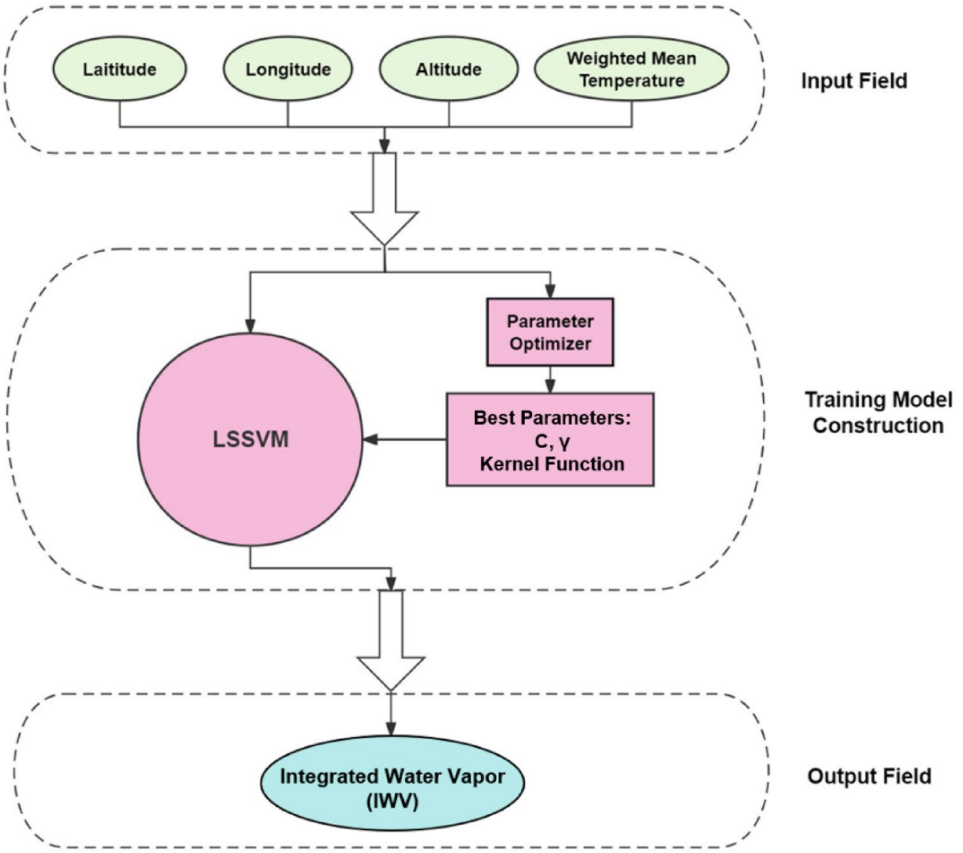

Fig.2 The flowchart of LSSVM in this research

While 187 stations were selected within the experimental region, some sites may be excluded from training at times due to considerations of data availability and quality assessments. The implemented estimation system covers a spatial extent ranging from the Upper Rhine Graben Region and its vicinity to a region approximately 350 km distant. Encompassing Germany, France, Switzerland, Luxembourg, as well as portions of Italy, the Netherlands, and Austria, totaling an area of 600,000 km^2. This selection was driven by the significance of climate change research in the Upper Rhine Graben Region, which holds critical implications for numerous European countries (Peters et al., 2007). Furthermore, the region exhibits diverse geographical characteristics, including plains, mountains, and low-lying areas. This study strives to develop a robust estimation model capable of accommodating diverse geographical scenarios. Subsequently, this study aims to apply our enhanced IWV estimation model to the designated region and evaluate its performance by comparing the results with radiosonde IWV and ERA5 IWV data.

Currently, several machine learning methods have been developed for estimating tropospheric delay. For instance, the tropospheric delay network (TropNet) model, based on deep learning techniques, has been applied to forecast ZWD. This model integrates data from the Geostationary Operational Environmental Satellite-R (GOES-R) series and the Global Forecast System (GFS) (Lu et al. 2023). Additionally, the backpropagation artificial neural network (BP-ANN) has been widely adopted in geoscientific research due to its adaptive, learning, and nonlinear capabilities. An adaptive regional 4-D water vapor density (WVD) model, utilizing the BP-ANN technique, was developed for vertical constraints in GNSS tropospheric tomography, based on sample data from ERA5 and GNSS meteorological observations from 2015 to 2019 (Zhang et al. 2024). In addition, there are several tropospheric product prediction models based on support vector machine (SVM) and artificial neural network (ANN) being proposed (Razin et al., 2022; Akar et al., 2023). For instance, the IWV prediction model based on single-site utilizes historical observation data from a specific GNSS site along with meteorological sensors installed at the site to forecast future IWV. However, this model is not applicable to areas lacking GNSS installations, leaving IWV variations in these regions unknown. Additionally, models based on SVM and ANN impose strict requirements for acquiring meteorological data, which poses challenges as most GNSS sites lack meteorological sensors. Therefore, this study proposes an IWV prediction model based on Least Squares Support Vector Machine (LSSVM), which can forecast unknown regions without relying on

stringent meteorological field data, achieving high-precision predictions with the assistance of ERA5 data. Compared to currently used SVM model, LSSVM establishes a linear decision function by introducing regularization parameters, sample error variables and other parameters. This not only enhances the fitting capability but also strengthens the model's generalization, while also accelerating the model training speed. The LSSVM algorithm is supported by a robust mathematical foundation, rendering it an effective tool for more stable IWV modeling. LSSVM has acquired considerable attention in diverse prediction tasks due to its robustness and efficiency (Mehrkanoon et al. 2012). Unlike conventional Support Vector Machines (SVM), which emphasize maximizing the margin between different classes, LSSVM prioritizes minimizing empirical risk, making it especially proficient in addressing regression issues. One of the significant advantages of LSSVM lies in its capability to effectively handle nonlinear relationships between input and output variables. Through the utilization of kernel functions, LSSVM can transform input data into high-dimensional feature spaces, facilitating the capture of complex patterns and structures with greater efficiency (Wang et al., 2007). This adaptability makes LSSVM well-suited for modeling complicated IWV systems characterized by nonlinearity. In the domain of predicting IWV values, LSSVM offers several distinct advantages. Firstly, IWV prediction necessitates capturing intricate relationships among various atmospheric parameters, which often exhibit nonlinear behaviors. By harnessing LSSVM's nonlinear modeling capabilities, we can better elucidate the complex dependencies inherent in atmospheric data. Additionally, LSSVM demonstrates robust performance even when confronted with relatively small datasets. This feature is particularly advantageous in meteorological science, where data accessibility may be limited, especially in remote or inaccessible regions. The capacity of LSSVM to furnish accurate predictions despite limited data can significantly advance our comprehension of IWV modeling and demonstrate immense value in various weather forecasting applications.

In our research, the application of LSSVM for IWV prediction offers a promising avenue to tackle the challenges linked with modeling high-accuracy IWV over large-scale areas. Coupled with the global distributed GNSS ZTD products sourced from NGL and the robust performance demonstrated by our enhanced IWV prediction model, as showed in our study, this approach holds remarkable potential to realize highly effective IWV prediction models on a global scale. Such improvements stand to enhance weather forecasting capabilities and contribute to the progress of atmospheric science research. The research begins by outlining the methodology of the proposed LSSVM scheme and the regional IWV estimation model. Following this, the experiment data is described. Subsequently, the results and validation procedures are given in detail. Finally, the conclusions drawn from the study are presented.

Methodology

LSSVM scheme design

SVM represent a category of supervised learning models extensively utilized in machine learning. These models are equipped with learning algorithms tailored to handle data analysis tasks encompassing both classification and regression (Vapnik et al. 1996). Grounded in statistical learning frameworks, SVM emerges as a powerful prediction method known for its robustness. Fundamentally, given a dataset comprising labeled examples belonging to two distinct categories, an SVM training algorithm constructs a model capable of effectively classifying new instances. By means of intricate mapping techniques, SVM projects training examples onto points within a multidimensional space, striving to maximize the margin separating the two categories. Consequently, new instances are projected onto this space and classified based on their position relative to the margin (Wang et al., 2005).

LSSVM offers an alternative approach within the domain of SVM, serving as supervised learning techniques applicable to both classification and regression scenarios. Unlike classical SVM, which tackles a convex quadratic programming (QP) problem, LSSVM obtains solutions through the resolution of a set of linear equations (Suykens et al., 1999). LSSVM belongs to the category of kernel-based learning methods, where traditional SVM formulations involve QP problems with inequality constraints. In contrast, LSSVM streamlines the problem by incorporating equality constraints. The training dataset consists of *M-th* input and output vectors:

$(\boldsymbol{x}_{train,i}, \boldsymbol{y}_{train,i}) \in \mathbf{V}_{train}, for \mathrm{i} = 1, 2, 3, \ldots, M,$

then, each input vector $\boldsymbol{x}_{train,i} : [x_{train,i1}, x_{train,i2}, \ldots, x_{train,ij}]$ is associated with corresponding output vectors $\boldsymbol{y}_{train,i} : [y_{train,i1}, y_{train,i2}, \ldots, y_{train,ik}]$. Here, the numbers of input and output parameters are present as j and k. Similarly, the test dataset:

$$\left(\boldsymbol{x}_{test,i}, \boldsymbol{y}_{test,i}\right) \in \mathbf{V}_{test}, for \mathrm{i} = 1, 2, 3, \ldots, N,$$

shares the same structure as the training dataset, but with *N-th* samples. When the new dataset represented by the vector $\boldsymbol{X}_{new}$ is introduced, its predicted output vector $\boldsymbol{Y}_{new}$ is determined by the training model, in accordance with the network structure of the designed scheme (Suykens et al., 1999). The LSSVM algorithm constructs a specialized function for each training sample $(\boldsymbol{x}_{train,i}, \boldsymbol{y}_{train,i})$, incorporating least-squares equation constraints to derive the corresponding Lagrange function equation. Optimal hyperparameters for the model are determined through cross-validation techniques. Subsequently, the optimal solution conditions and the Radial Basis Function (RBF) kernel function K are combined to form the LSSVM regression model (Li et al. 2021). The aforementioned procedures are illustrated by the following equations (Suykens et al., 1999):

The basic LSSVM is introduced by the problem and minimize the error function $\mathcal{F}_3(\omega, b, e)$ as:

$$\min_{\omega,b,e} \mathcal{F}_3(\omega, b, e) = \frac{1}{2}\omega^T\omega + C\frac{1}{2}\sum_{k=1}^{N} e_k^2 \tag{1}$$

where ω is the weight vector, b presents the model parameter, C denotes to the regularization constant and e_k refers to the error vector of the model. Then, this equation subjects to the equality constraints

$$\boldsymbol{y}_{train,k}\left[\omega^T\varphi\left(\boldsymbol{x}_{train,k}\right) + b\right] = 1 - e_k, k = 1, \ldots, N \tag{2}$$

where φ represents the mapping function and we design the Lagrangian equation:

$$L_3(\omega, b, e; \alpha) = \mathcal{F}_3(\omega, b, e) - \sum_{k=1}^{N} \alpha_k \left\{\boldsymbol{y}_{train,k}\left[\omega^T\varphi\left(\boldsymbol{x}_{train,k}\right) + b\right] - 1 + e_k\right\} \tag{3}$$

where α_k are Lagrange multipliers (which can be positive or negative) based on the equation between $\boldsymbol{x}_{train,k}$ and $\boldsymbol{y}_{train,k}$ as:

$$\boldsymbol{y}_{train,k} = \omega^T\varphi\left(\boldsymbol{x}_{train,k}\right) + b + e \tag{4}$$

finally, we introduce the kernel function:

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \mathrm{e}^{\frac{-\left|\left|x_i - x_j\right|\right|^2}{2\sigma^2}} with \sigma^2 = \frac{1}{2\gamma} \tag{5}$$

$$\boldsymbol{y}_{train,k} = \sum_{k=1}^{N} \alpha_k \mathrm{K}\left(\boldsymbol{x}_{train,k}, \boldsymbol{x}_{train,j}\right) + b \tag{6}$$

where γ in Eq. (5) is a parameter of the kernel function, which implicitly determines the distribution of data mapped to the new feature space. σ denotes the standard deviation of Gaussian distribution, which is determined through the cross-validation optimization algorithm. Therefore, when γ is too large, that is, σ is too small, the corresponding Gaussian distribution will become narrower and the support vectors will become fewer, causing the model to perform well only on samples near the support vectors and poorly classify unknown samples. On the contrary, when γ is too small, the smoothing effect will be too large and it will not be possible to achieve particularly high accuracy on the training dataset.

Based on the LSSVM algorithm, this study develops a customized approach to construct the IWV model using the provided input and output data. The essence of this approach lies in leveraging the LSSVM algorithm to delineate the functional correlation between station position and meteorological data in relation to IWV, as depicted in Fig. 2. The input variables consist of GNSS station longitude, latitude, altitude, and the weighted mean temperature from the station height. Subsequently, the output comprises the estimated GNSS IWV values corresponding to these positions. The inclusion of temperature parameters in the input field is crucial to this approach. Regarding the input data for the LSSVM algorithm, the position information, including latitude (λ), longitude (φ) and altitude (H) is readily available in the GNSS RINEX observation file. However, the weighted mean temperature is not directly provided. Therefore, this study employs interpolation to derive this parameter from the corresponding ERA5 pressure level data.

In the establishment of LSSVM training models, the selection of parameters such as the regularization constant C, the variance σ of the kernel function, and the kernel function $\mathrm{K}\left(x_i, x_j\right)$ plays a pivotal role in determining the accuracy of model and generalization capability. These parameters govern the trade-off between model complexity and generalization performance. Thus, their optimal values are crucial for achieving high predictive accuracy and robustness across different datasets. To address this optimality issue, our study adopts a rigorous approach by employing cross-validation techniques on randomly partitioned training and testing samples. Through multiple iterations of cross-validation, we systematically explore the parameter space to identify the combination of C, σ, and the kernel function that yields the lowest errors on the given dataset. This process ensures that the model is fine-tuned to capture the underlying relationships between input and output parameters effectively. Furthermore, by utilizing advanced optimization algorithms and statistical techniques, we aim to refine the parameter selection process and enhance the model's precision and applicability. By selecting the most appropriate parameters for each training dataset, our methodology improves the model predictive accuracy and its ability to generalize to unseen data. This meticulous approach not only enhances the reliability of our estimation model but also contributes to the advancement of predictive modeling techniques in the field.

Retrieval of the weighted mean temperature

The methodology of obtaining the weighted mean temperature is outlined as follows (Bevis et al. 1992):

$$T_m = \frac{\sum_{H_s}^{H_{top}} \frac{e_j}{T_j} \Delta H_j}{\sum_{H_s}^{H_{top}} \frac{e_j}{T_j^2} \Delta H_j} \tag{7}$$

where T_m is the weighted mean temperature, e and T denote the water vapor pressure (in hPa) and temperature (in K), respectively, spanning from the altitude of the GNSS station (H_s) to the upper limit of the troposphere (H_{top}). Consequently, T_m (in K) can be computed using the ERA5 pressure level product, which provides data on an hourly basis, with a spatial resolution of 0.25°×0.25°, and covers 37 vertical pressure levels.

Ideally, each GNSS station would be equipped with meteorological sensors to obtain temperature, pressure, and other meteorological data at its location. However, setting up such infrastructure requires significant manpower and resources. Consequently, not every GNSS station is currently equipped with meteorological sensors. Therefore, this research opts to utilize ERA5 grid data for calculations instead. However, ERA5 products are structured in a grid format, which may not align precisely with the locations of all GNSS stations. Thus, when GNSS stations do not coincide with grid points, it becomes necessary to interpolate ERA5 data from the nearest grid points to the respective GNSS station positions, accounting for both horizontal and vertical dimensions. In this study, our focus is specifically on pressure and temperature variables, which can be interpolated using the same bilinear interpolation method for horizontal interpolation. Bilinear interpolation is used because it effectively handles the two-dimensional horizontal data, considering the values at the four nearest ERA5 grid points to provide a smooth and accurate estimation. The study commences with the linear vertical interpolation method, as temperature or pressure and height exhibit a linear relationship, enabling the straightforward computation of pressure and temperature at the GNSS station altitude based on the nearest vertical levels. In some cases, the altitude of GNSS stations is below than the surface height of the ERA5 reanalysis data, attributable to the following factors: First, discrepancies in data sources exist; the surface elevation in ERA5 reanalysis data is primarily derived from digital elevation model (DEM) and meteorological station observations, which may exhibit variations in resolution and accuracy compared to the actual measurements obtained from GNSS stations. Second, there are considerations regarding data updates and temporal lags: GNSS measurement data are typically updated in real-time, whereas ERA5 reanalysis data are generated from historical observations and model predictions, which may introduce inherent temporal lags and variability. In order to remain the integrity of input parameter precision and the number of training samples is sufficient, we exclude some GNSS stations whose on-site measured height is lower than the corresponding ERA5 surface elevation. The vertical interpolated procedures of temperature are described below (Jade et al., 2008):

$$T_j^i = \frac{\omega_j^{i,i-1}}{\omega_j^{i,i-1}+\omega_j^{i,i+1}} T_j^{i,i-1} + \frac{\omega_j^{i,i+1}}{\omega_j^{i,i-1}+\omega_j^{i,i+1}} T_j^{i,i+1} \tag{8}$$

with weighting factors $\omega_j^{i,i-1} and \omega_j^{i,i+1}$

$$\omega_j^{i,i-1} = \frac{1}{\left(H_j^{i-1}-H\right)^2} and \omega_j^{i,i+1} = \frac{1}{\left(H_j^{i+1}-H\right)^2} \tag{9}$$

with the intermediate parameters $T_j^{i,i-1} and T_j^{i,i+1}$

$$T_j^{i,i-1} = T_j^{i-1}\left[1+\left(\frac{8.419\times10^{-5}\left(H_j^{i-1}-H\right)}{\left(T_j^{i-1}\right)^{0.1902884}}\right)\right]^{5.255303}$$

$$T_j^{i,i+1} = T_j^{i+1}\left[1+\left(\frac{8.419\times10^{-5}\left(H_j^{i+1}-H\right)}{\left(T_j^{i+1}\right)^{0.1902884}}\right)\right]^{5.255303} \tag{10}$$

where H_j^{i-1} and H_j^{i+1} are the heights at two pressure layers corresponding to the pressure level P_j^{i-1} and P_j^{i+1} respectively. Similarly, T_j^{i-1} and T_j^{i+1} denote the temperatures at two pressure layers corresponding to the pressure level P_j^{i-1} and P_j^{i+1} respectively. T_j^i is the interpolated temperature at the grid point corresponding to the height H of the GNSS station.

Following the calculation of temperature at four grid points corresponding to the GNSS station height, the temperature values at the GNSS station positions are acquired through the employment of the bilinear interpolation method for horizontal interpolation (Zhang et al. 2023):

$$T = \sum w_j^i T_j^i = w_1^i T_1^i + w_2^i T_2^i + w_3^i T_3^i + w_4^i T_4^i \tag{11}$$

where

$$w_j^i = \frac{w_{g_j}^{\ i}}{w_{g_1}^{\ i}+w_{g_2}^{\ i}+w_{g_3}^{\ i}+w_{g_4}^{\ i}} for j = 1,2,3,4 \tag{12}$$

with $w_{g_j}^{\ i}$ is the weighting coefficient for the grid points

$$w_{g_j}^{\ i} = \left(R\mu_j^i\right)^{-1} \tag{13}$$

$$\cos\mu_j^i = \sin\varphi_j^i \sin\varphi + \cos\varphi_j^i \cos\varphi \cos\left(\lambda_j^i - \lambda\right) \tag{14}$$

where R is the mean radius of the Earth, μ_j^i is the angular distance between the four grid points $(\lambda_j^i, \varphi_j^i)$ and the position of the GNSS station (λ, φ). Subsequently, T represents the interpolated temperature at the point (λ, φ, *H*). These interpolation equations are also applicable for pressure (P) interpolation. Referring to the Equations from (7) to (14), we can compute the weighted mean temperature at the GNSS station's position and altitude.

Retrieval of IWV from ZTD

In the context of LSSVM output parameter GNSS IWV, this research opts for individual computations. In most cases, the post-processing software's tropospheric wet part product is unreliable. Consequently, this study exclusively utilizes their ZTD product to derive IWV, employing the following procedures (Saastamoinen 1972; Davis et al., 1985; Yuan et al. 2022):

$$ZHD = 2.2768\frac{P_s}{1-2.66\times10^{-3}\bullet\cos(2\varphi_s)-2.8\times10^{-7}H_s} \tag{15}$$

where P_s is GNSS station pressure (in hPa) and φ_s and H_s are latitude (in rad) and orthometric height (in m) of the GNSS station and ZWD can be obtained:

$$ZWD = ZTD - ZHD \tag{16}$$

introducing the relationship between ZWD and IWV:

$$IWV = Q \bullet ZWD \tag{17}$$

$$Q = \frac{10^6}{R_v\bullet\left[k_2\prime+\frac{k_3}{T_m}\right]} \tag{18}$$

where Q is the conversion factor and $R_v = 461.522Jkg^{-1}K^{-1}$ (Kestin et al. 1984) is the specific gas constant for water vapor. The parameters $k_2\prime = 22.1KhPa^{-1}$ and $k_3 = 373900K^2hPa^{-1}$ are atmospheric refractivity constants (Bevis et al. 1994).

Retrieval of validation data: ERA5 IWV and radiosonde IWV

The computational approach for ERA5 IWV is outlined. Employing both the ZTD ray tracing model and integral methods, IWV values are computed for GNSS stations using ERA5 data. Initially, the ERA5 dataset comprises hourly geopotential height (in m), temperature (in K), and relative humidity (in %) across 37 pressure levels. To provide a succinct overview of the procedures: The initial step involves converting geopotential height to ellipsoid height based on the WGS84 ellipsoid. Subsequently, meteorological interpolation work is conducted following Eqs from (7)–(14), facilitating the acquisition of meteorological parameters for GNSS stations. Next, the integration of ZTDs across all layers, from the GNSS station height to the tropopause, is performed. For each layer, the processing procedures are meticulously delineated as follows (Jiang et al. 2023):

$$ZTD = 10^{-6}\int_{H_{station}}^{H_{top}} NdH = 10^{-6}\sum_{H_{station}}^{H_{top}} N_i\Delta H_i \tag{19}$$

$$N = k_1\left(\frac{P-E}{T}\right) + k_2\left(\frac{E}{T}\right) + k_3\left(\frac{E}{T^2}\right) \tag{20}$$

where $k_1 = (77.604 \pm 0.014)\ KhPa^{-1}$, $k_2 = (64.79 \pm 0.08)\ KhPa^{-1}$ and $k_3 = (377{,}600 \pm 400)K^2hPa^{-1}$. N, H_{top}, and $H_{station}$ denote the atmospheric refraction, height of the top level, and height of the GNSS station, respectively. To introduce the partial water vapor pressure E (in hPa):

$$E = 0.06112 \bullet RH \bullet e^{\frac{22.46\bullet(T-273.15)}{272.62+T-273.15}} \tag{21}$$

with RH (in %) and T (in K) denote the relative humidity and temperature, respectively. After calculating ERA5 ZTD with Eqs. (19)–(21), the estimation of ERA5 IWV is conducted following the equations from Eqs. (15)–(18).

Next, the method for calculating IWV based on Radiosonde data is presented. Similar to ERA5 layer data, Radiosonde data is also provided in profiles, typically containing temperature (in °C), dew point temperature (in °C), pressure (in Pa), geopotential height (in m), wind direction (in degrees from north), and wind speed (in m/s) for each profile. Figure 3 depicts the pressure and temperature profiles at Station BEM00006458 for day of year (DOY) 252, 2022, serving as an illustrative example. The graph effectively showcases the temperature fluctuations across various pressure levels. The bold red line represents the vertical temperature profile, while the bold green line denotes the vertical dew point temperature. The vertical axis indicates air pressure in hPa, with horizontal lines on the graph representing lines of constant pressure.

The integral method is employed to compute Radiosonde IWV across the observational region. The radiosonde profiles are quality-controlled by removing empty values. If the resulting cleaned radiosonde profiles contain fewer than 5 layers, the data for that epoch is discarded. The calculation of Radiosonde IWV from the radiosonde profiles is outlined as follows (Namaoui et al. 2021):

$$IWV = \int_{H_s}^{H_{top}} \rho_w dh = \sum_{i=1}^{n} \rho_{d,i} r_i \Delta h_i \tag{22}$$

where ρ_w is the water vapor density and n denotes the number of layers. $\rho_{d,i}$, r_i and Δh_i refer to the dry air density, the mixing ratio, and the altitude height difference of each layer. For each layer, the parameters are expressed as:

$$\rho_d = \frac{M_d}{R}\frac{P}{T}, r = 0.622\frac{e}{P} \tag{23}$$

with the molar mass od dry air $M_d = 0.0289644kgmol^{-1}$ and the universal gas constant for air $R = 8.31432Jmol^{-1}K^{-1}$. T is the temperature (in K) and P is the pressure (in Pa). In addition, the partial water vapor pressure e (in Pa) is introduced based on the saturation vapor pressure e_s (in Pa) as follows:

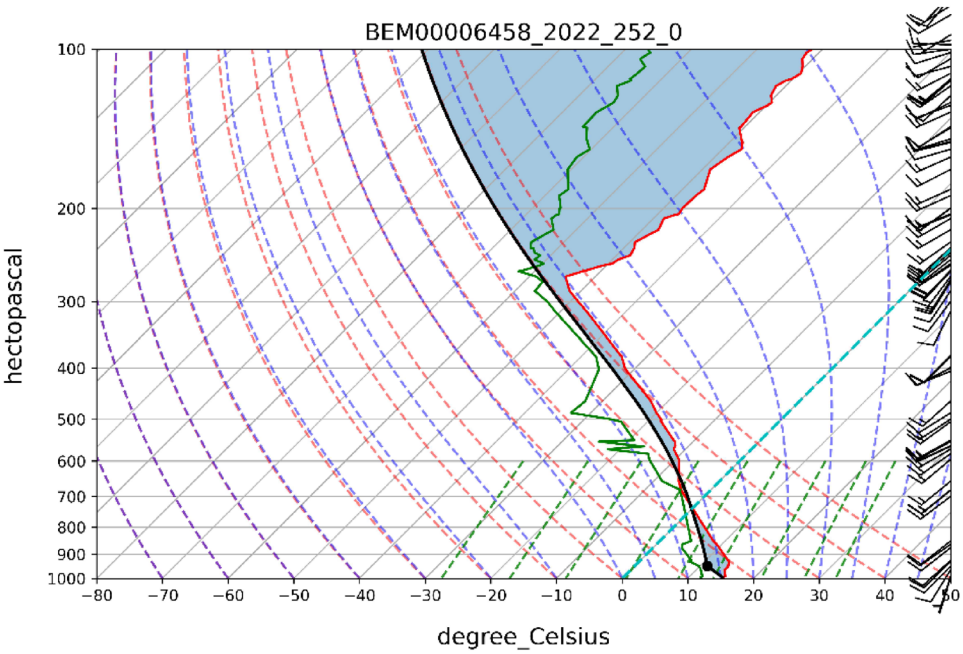

Fig. 3 Radiosonde profiles for station BEM00006458 at 00:00 on day 252, 2022. The red line indicates the temperature, the green solid line represents the dew point temperature, and the black line shows the pressure and temperature at the lifting condensation level. The dotted lines denote the isotherms. The wind vane on the right illustrates the wind speed and direction, with 0 degrees representing due north, 180 degrees due south, and increasing clockwise. (Radiosonde data can be downloaded from: https://climatedataguide.ucar.edu/climate-data/integrated-global-radiosonde-archive-igra/)

$$e = \frac{RH}{100} e_s \tag{24}$$

where RH is relative humidity (in %) and e_s are calculated by the temperature T (in °C) and the dew point temperature T_d (in °C):

$$e_s = 6.1094 \bullet \mathrm{e}^{\left(\frac{17.625 \bullet T}{T+243.04}\right)} \tag{25}$$

$$RH = 100 \bullet \left[\frac{\mathrm{e}^{\left(\frac{17.625 \bullet T_d}{T_d+243.04}\right)}}{\mathrm{e}^{\left(\frac{17.625 \bullet T}{T+243.04}\right)}}\right] \tag{26}$$

Accuracy evaluation

The evaluation of the model's accuracy is then presented. Owing to the model training based on cross-validation, the evaluation of model performance is conducted through self-assessment leave-one-out (LOO) method, utilizing GNSS IWV at every time epoch. The LOO cross-validation, also known as n-fold cross-validation, where n represents the number of samples. This methodology involves conducting n iterations to evaluate the predictive performance of the model at each GNSS station. In each iteration, one station is reserved as the test dataset, while the remaining n-1 samples serve as training data, which is showed in Fig. 4. This approach offers the advantage of maximizing the utilization of training samples in each iteration, while also providing deterministic predictions, ensuring consistent outcomes across iterations.

Subsequent validation involves testing the IWV estimation model against external sources, such as radiosonde IWV, to demonstrate its reliability and high accuracy. The validation IWV is computed from radiosonde data, serving as the reference value due to its high precision derived from superior sensor quality and calibration capabilities. Our proposed LSSVM model is then applied to the positions of

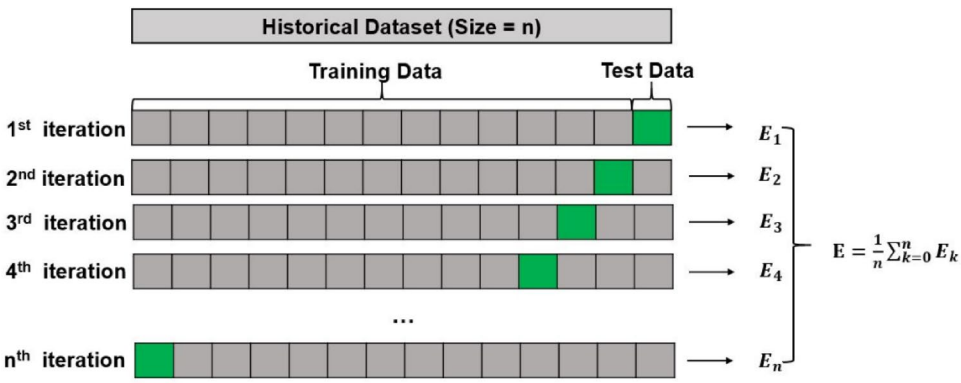

Fig. 4 The flowchart of Leave-one-out cross validation and E_n is RMSE at the n-th iteration

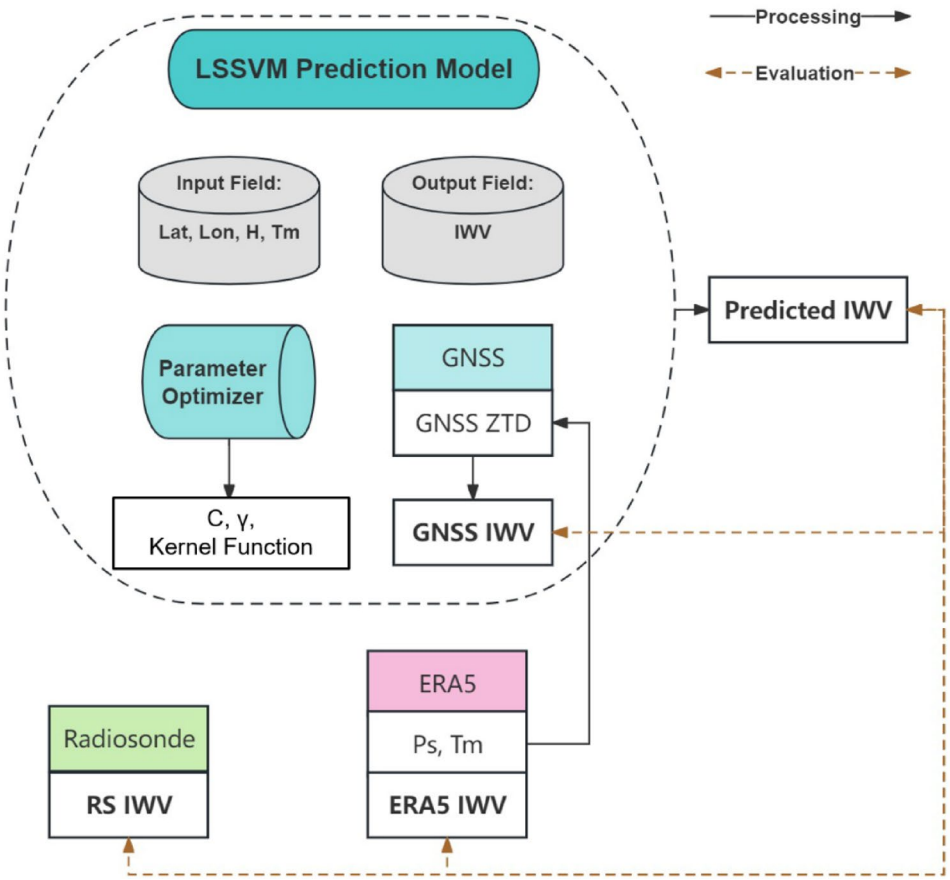

Fig. 5 The flowchart of the applied methodology

these radiosonde stations, with the predictive IWV compared against actual radiosonde IWV to establish the accuracy. Furthermore, ERA5 IWV validation data is incorporated for comparison with our model estimations, illustrating the model's enhancements in IWV estimation compared to the traditional method when the GNSS measurements or radiosonde stations are absent, particularly in large-scale regional IWV estimation capabilities. In this study, linear regression is employed to validate various derived IWV datasets. Three comparison groups are considered, involving the comparison of predicted IWV with GNSS IWV, predicted IWV with ERA5 IWV, and predicted IWV with radiosonde IWV. The following formulas are utilized for statistical analysis:

$$MBE = \frac{\sum_{i=1}^{N}(IWV_1 - IWV_2)}{N} \tag{27}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(IWV_1 - IWV_2)^2}{N}} \tag{28}$$

$$R = \frac{\sum_{i=1}^{N}\left(IWV_1 - \overline{IWV_1}\right)\left(IWV_2 - \overline{IWV_2}\right)}{\sqrt{\sum_{i=1}^{N}\left(IWV_1 - \overline{IWV_1}\right)^2 \sum_{i=1}^{N}\left(IWV_2 - \overline{IWV_2}\right)^2}} \tag{29}$$

let N represent the total number of IWV values, where $\overline{IWV_1 and IWV_2}$ denote two derived IWV datasets, and $\overline{IWV_1}$, $\overline{IWV_2}$ represent their respective mean values. Subsequently, our proposed model is evaluated by comparing the Mean Bias Error (MBE), Root Mean Square Error (RMSE), and Correlation (R) between the two derived IWV datasets.

The final methodology of the paper is summarized in the flowchart depicted in Fig. 5.

Data description

GNSS data processing and IWV retrieval

The GNSS technique has emerged to a promising approach for acquiring IWV, offering high temporal resolution IWV data across diverse weather conditions. Numerous studies have evaluated IWV accuracy against various reference datasets worldwide, consistently demonstrating accuracy levels typically below 3 mm (Prasad et al., 2009; Baba et al., 2022). This study utilized one year of GNSS receiver measurements from approximately 150 stations in Upper Rhine Graben Region (January 2022 to December 2022). Leveraging NGL

Table 1 GNSS observation postprocessing parameters

Parameter	Value
GNSS Observation Period	2022 001–2022 365
GNSS Processing Software	GipsyX-1.0
Trop Mapping Function	VMF1
A Prior Tropospheric Delay Model	VMF1/NWM grid
Horizontal Gradient Model	Chen and Herring model
Processing Mode	PPP
Ionosphere Effects	None
Cut-off Angle	7°
Sample Interval	300 s
Other Inputs	JPL orbits/clocks/IONEX; VMF1/NWM grid; Chalmers/FES2004

products,[1] the processing steps for GNSS data are outlined in the Table 1: the GNSS receiver data underwent post-processing using GipsyX-1.0 software to compute tropospheric delays by using Ionospheric delay free data. Parameters configured during GipsyX-1.0 utilization included: Vienna Mapping Function 1 (VMF1) for mapping function, gradient model following Chen and Herring (Chen and Herring, 1997), cutoff angle set at 7°, GNSS constellation limited to GPS, and Precise Point Positioning (PPP) technique employed. The tropospheric delay comprises hydrostatic (dry) and non-hydrostatic (wet) delay components. The VMF1 dry grid model are used to estimate the hydrostatic delay, subsequently subtracted from the tropospheric delay to yield the wet delay. Finally, ZWD was converted into IWV with introducing the weighted mean temperature. In this research, ZTD Product from NGL are utilized.

Validation data: ERA5 data and radiosonde data

The European Centre for Medium-Range Weather Forecasts (ECMWF)[2] is dedicated to providing accurate and timely weather forecasts, climate monitoring, and environmental research. Through its sophisticated modeling systems and vast array of observational data, ECMWF delivers valuable insights into global weather patterns and climate dynamics. This research utilizes the ERA5 dataset for validation, which offers comprehensive atmospheric reanalysis data at high spatial and temporal resolutions. The ERA5 dataset spans from 1979 to present and provides detailed information on various meteorological parameters, including geopotential, relative humidity, and temperature, among others. The dataset is available with a grid resolution of 0.25° in both latitude and longitude, ensuring fine-scale representation of atmospheric conditions. Specifically, our study utilizes ERA5 37 pressure levels data, covering the longitude range of 4–12°E and the latitude range of 45–51°N. The data encompasses the entirety of the year 2022, with a temporal resolution of 6 h. This comprehensive dataset from ECMWF serves as a crucial resource for the research, facilitating the validation and analysis of proposed IWV estimation model within the specified region and time frame.

In this research, we intend to utilize data sourced from radiosonde, accessed from the Integrated Global Radiosonde Archive (IGRA)[3] for the entirety of the year 2022 (Durre et al. 2016). We selected 9 radiosonde stations within the research domain, where radiosonde balloons were launched twice daily from each station. This study shows the basic information of these stations in Table 2 and Fig. 6. To ensure the reliability and integrity of the dataset, rigorous quality control measures were applied to the radiosonde data. This included scrutiny of data validity and the completeness of stratified data layers. By systematically eliminating unsatisfactory data points, this study processed the remaining radiosonde balloon data to derive the corresponding IWV values. The specific methodology employed for IWV calculation is comprehensively detailed in the methodology section. These meticulously processed radiosonde data constitute a vital

Table 2 The basic information of selected radiosonde stations for validation: the position, the altitude, and the launched city

Station Name	Latitude: (°)	Longitude: (°)	Altitude: (m)	Region
GMM00010739	48.8333	9.2000	314.0	STUTTGART
BEM00006458	50.7456	4.7633	112.8	BEAUVECHAIN
GMM00010548	50.5617	10.3772	450.0	MEININGEN
GMM00010618	49.6928	7.3264	376.0	IDAR-OBERSTEIN
GMM00010771	49.4283	11.9022	417.0	KUEMMERSBRUCK
GMM00010868	48.2442	11.5525	484.0	OBERSCHLEISSHEIM
GMM00010954	47.8342	10.8667	756.0	ALTENSTADT
GMM00010962	47.8019	11.0119	977.0	HOHENPEISSENBERG
SZM00006610	46.8117	6.9425	490.0	PAYERNE

[1] CONTACT: Geoffrey Blewitt; gblewitt@unr.edu; http://geodesy.unr.edu

[2] Data can be downloaded from https://www.ecmwf.int/

[3] Data can be downloaded from https://climatedataguide.ucar.edu/climate-data/integrated-global-radiosonde-archive-igra/

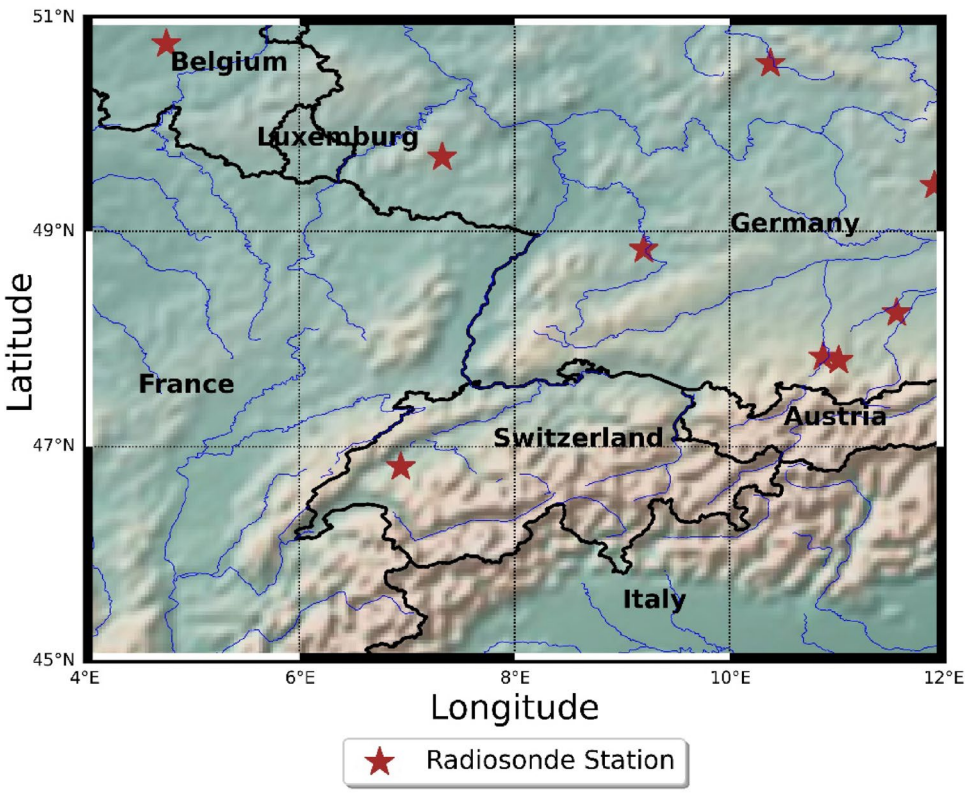

Fig. 6 The distribution map includes 9 radiosonde stations used for validation

benchmark for evaluating the accuracy and efficacy of our novel modeling approach.

Test case

The Upper Rhine Graben Region holds significant importance in atmospheric research due to its unique topographical and meteorological characteristics. Several notable research papers have focused on this region, shedding light on its climatic characteristics and atmospheric dynamics. For instance, Timar-Geng et al. (2006) conducted a comprehensive analysis of temperature trends in the region, highlighting the region's susceptibility to heatwaves and temperature extremes. Additionally, Thierion et al. (2012) explored the water balance and precipitation patterns in this area, revealing significant alterations in rainfall distribution. The Upper Rhine Graben Region, situated in Central Europe, provides a diverse landscape comprising plains, mountains, and low-lying areas. This geographical diversity contributes to the region's dynamic weather patterns, which are influenced by its proximity to major water bodies such as the Rhine River and the Alps. DEM[4] of this region is showed in Fig. 7. As a result, this region serves as a focal point for numerous research endeavors aimed at understanding weather dynamics, climate variability, and hydrological processes.

Research conducted within this region is necessary for several reasons. The area has diverse geographical features, as illustrated in Fig. 7, ranging from plains to mountains, with altitudes exceeding 4,000 m. In addition, its temperate climate showcases distinct seasonal variations. According to data from the German Weather Service (DWD),[5] summers are typically warm, with average temperatures ranging from 17 to 25 °C, while winters are cold, often accompanied by snowfall, especially at higher elevations, with average temperatures from −1 to 5 °C. Water vapor is the important factor in shaping the atmospheric conditions of the Upper Rhine Graben Region. Accurate estimation models of IWV are therefore advantageous for studying weather and precipitation patterns within this area.

Results and analysis

In this section, we provide a detailed compilation and discussion of the comprehensive findings derived from our analysis. Initially, we present the accuracy assessment results of the LSSVM prediction model through leave-one-out (LOO) cross validation and validation against ERA5 IWV data. The comparative analysis is conducted across various epochs, highlighting the model's performance across different station

[4] Data can be downloaded from https://www.usgs.gov/

[5] Data can be downloaded from https://werdis.dwd.de/

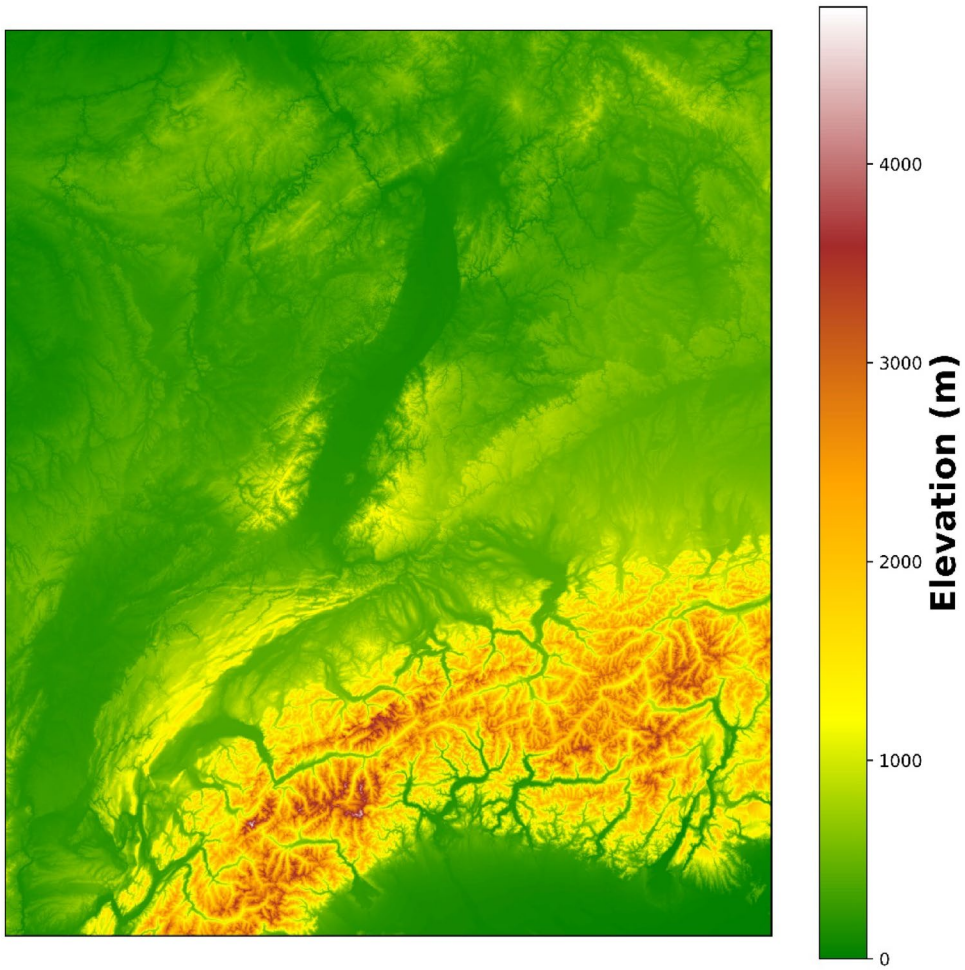

Fig. 7 DEM of the Upper Rhine Graben region, spanning geographical latitudes from 45°N to 51°N and longitudes from 4°E to 12°E

series. Subsequently, we validate our LSSVM model against radiosonde IWV data and present the results for each station in the form of time series plots. This rigorous validation process offers insights into the model's performance and robustness across different temporal and spatial scales. Furthermore, we delve into the inner parameters of the LSSVM model, exploring the relationship between parameter selection and model performance. By elucidating the impact of parameter choices on model performance, we provide valuable insights for refining the model and optimizing its predictive capabilities. Last, we utilize our proposed methodology to generate regional maps depicting short-term or long-term variations in IWV, offering a comprehensive visualization of IWV distribution across the study area. This final step serves to synthesize our findings and provide a holistic understanding of IWV patterns and variability within the region of interest.

Assessment of the LSSVM prediction model

In this research, the machine learning method is employed to predict IWV using the LSSVM algorithm. In our case, the input parameters considered include latitude, longitude, altitude, and the weighted mean temperature, all are easily accessible. Meanwhile, maintaining high precision in the input data is crucial for optimal model performance. Since latitude, longitude, and altitude are provided directly, the accuracy of the weighted mean temperature significantly influences model performance for several reasons. First, the robustness of the model depends on the accuracy of the input data. The model is designed to learn the mapping patterns between input parameters and the output parameter, and any imprecise input can mislead the regression process. Second, the weighted mean temperature is vital for converting ZWD to IWV. Any inaccuracies in the weighted mean temperature can propagate through the model, ultimately impacting the accuracy of IWV estimates. Therefore, ensuring the precision of the weighted mean temperature is essential for the model to perform reliably and to mitigate potential errors arising from imprecise temperature data during training. The GNSS IWV serves as the output for our model, enabling its implementation across various locations worldwide where GNSS stations are available. Given the global distribution of GNSS stations, the potential for designing regional LSSVM

prediction models across different geographical regions is highlighted. This wide coverage allows for the development of our predictive model suited to specific localities, thereby enhancing the model's versatility and applicability across diverse geographic contexts. The subsequent content provides an in-depth analysis of the performance of our proposed model.

Given the seasonal dependency of water vapor activity, the model performance is evaluated based on summer and winter. As depicted in Fig. 8, winter displays three days representing water vapor inactivity, while summer exhibits three days active of water vapor activity. All datasets undergo cross-validation procedures to ensure robustness and reliability in analysis. The prediction results from the proposed model are presented, with Fig. 8 illustrating the model's estimation (Predicted IWV) compared to GNSS IWV (Reference IWV). This comparative analysis provides valuable insights into the effectiveness and accuracy of the predictive model across varying atmospheric conditions. In this figure, the horizontal axis represents the GNSS station number, while the vertical axis denotes IWV with the unit of millimeters (mm). The green line is the predicted IWV from our proposed estimation model and the black line is GNSS IWV as reference values. The prediction results for different days of the year (DOY) reveal variability in the number of stations due to some GNSS stations lacking troposphere products, including ZTD, on specific days. Consequently, the size of the training model dataset is contingent upon the availability of GNSS products across different epochs.

It is observed that the active IWV exhibits an average variation of approximately 25 mm, whereas the inactive IWV varies by approximately 12 mm on average. This discrepancy indicates distinct water vapor variation trends between active and inactive periods, leading to different performances of the prediction model across the two periods. Nevertheless, whether depicting active or inactive periods, the predicted IWV closely aligns with reference IWV, accurately reflecting IWV variation. However, some stations exhibit relatively large differences which are considered as outliers in our research. Consequently, we introduced a statistical analysis as a validation

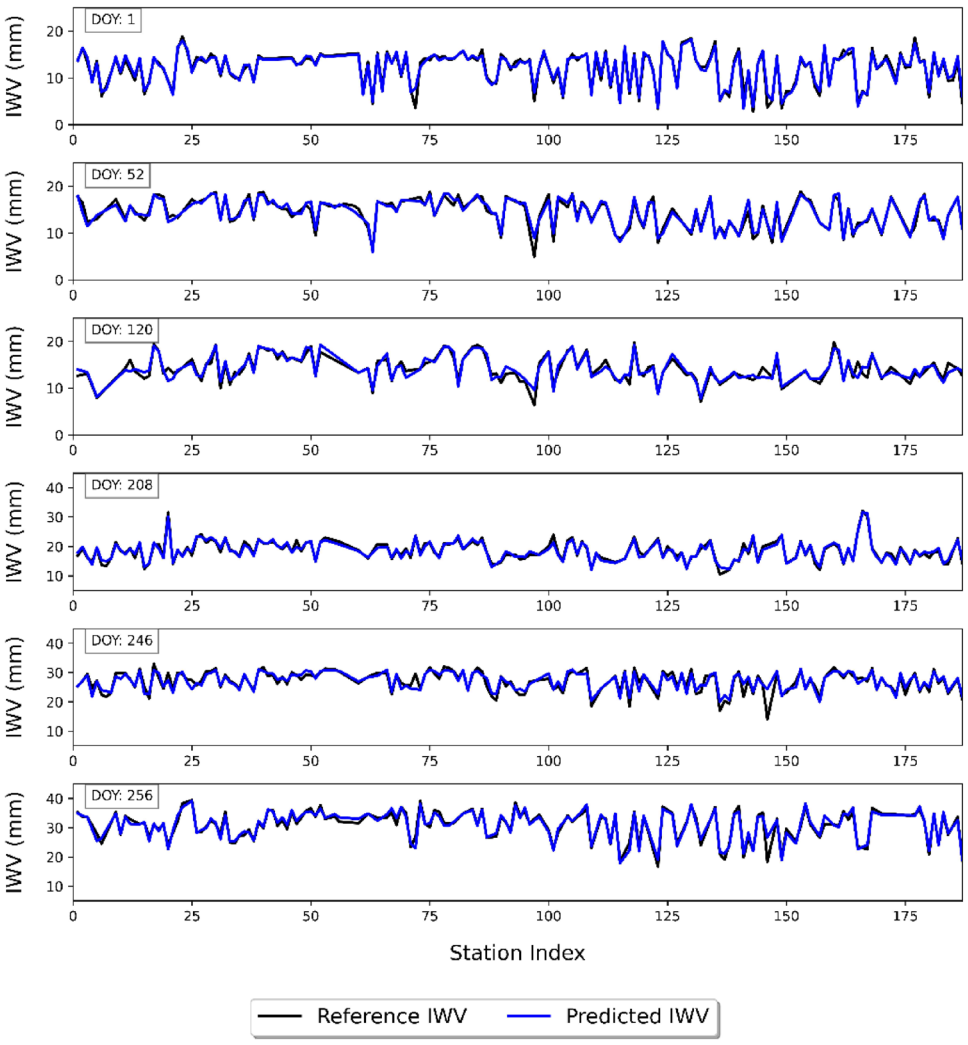

Fig. 8 The comparison between Predicted IWV and Reference IWV is conducted at specific epochs: 06:00 h on Day 1, Day 52, and Day 120 of the year, representing winter without significant water vapor activity (inactive period) with IWV range from 0 to 20 mm. Conversely, at specific epochs: 06:00 h on Days 208, 246, and 256, corresponding to summer with significant water vapor activity (active period) with IWV range from 5 to 45 mm

step to address these outliers and ensure the reliability of our predictions. In order to identify and analyze outliers within our dataset, we employed the 3-sigma rule, a widely utilized statistical method for outlier detection. This approach is based on the principle that data points lying beyond three standard deviations from the mean are considered as outliers. Mathematically, outliers can be detected using the following formula:

$$UpperLimit = \mu + 3\sigma and LowerLimit = \mu - 3\sigma \quad (30)$$

where μ is the mean value of dataset and σ denotes the standard deviation of dataset, Then, data points falling outside these upper and lower limits are flagged as outliers. Subsequently, the values of these outliers are visualized using box plots in Fig. 9, providing a comprehensive overview of their distribution within the dataset. We conducted a comparative analysis by computing the difference between the model predicted IWV and the actual IWV, followed by outlier detection. Our findings reveal that the predictive model does not exhibit consistent overestimation or underestimation. Instead, outliers are observed on both limits of the boxplot, consistent with our expectations. Additionally, we use the Valid Rate (VR) to evaluate the proportion of the number of outliers ($N_{outlier}$) relative to the total number of samples (N_{all}) according to Eq. (31). A higher VR value indicates greater model applicability and a lower proportion of outliers. However, the proportion of outlier stations remains below 5% of the total sample data, indicating that the predictive model generally maintains superior predictive capability across different time periods. Table 3 provides a summary of the accuracy of predictive model, including R, MBE, RMSE, and Valid Rate. It is noteworthy that even within the active period, certain days can exhibit relatively low and stable water vapor conditions. For example, on DOY 208,

Table 3 The accuracy of our prediction model on selected 6 epochs

Period	DOY	R	MBE (in mm)	RMSE (in mm)	VR (%)
Inactive	1	0.96	−0.05	0.76	97.6
	52	0.95	−0.04	0.81	97.9
	120	0.93	−0.07	0.96	97.0
Active	208	0.95	0.03	0.81	98.2
	246	0.84	−0.05	1.56	97.6
	256	0.96	−0.08	1.26	98.0

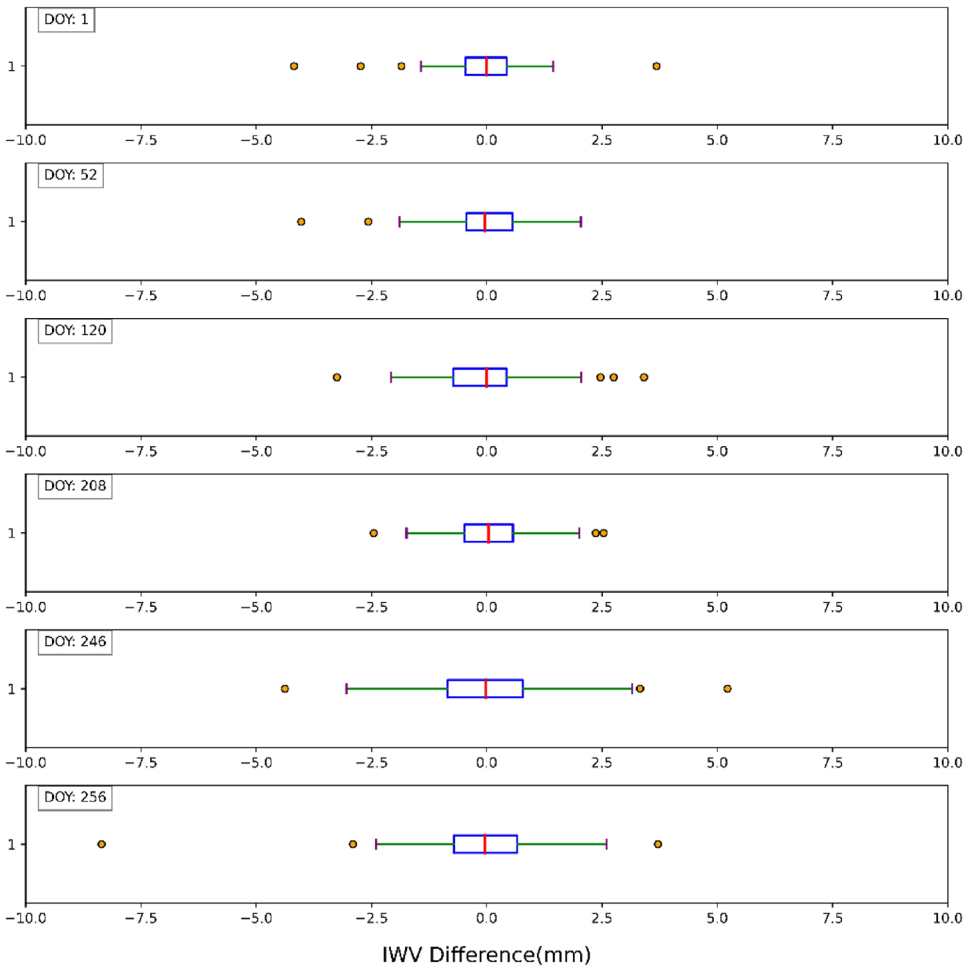

Fig. 9 The boxplots of IWV difference between predicted IWV and reference IWV

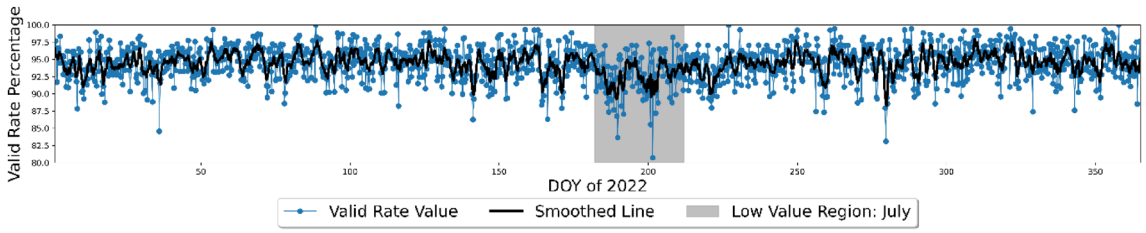

Fig. 10 The VR values for the entire year of 2022 are depicted in blue dots, representing individual data points. The black line is a fitted and smoothed line, which clearly illustrates the overall trend. The grey shaded area highlights the month of July, indicating a period of low valid rate throughout the year

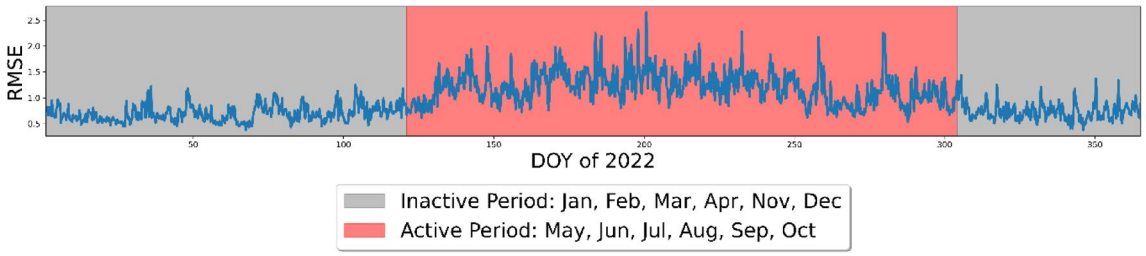

Fig. 11 The RMSEs for the entire year of 2022 are depicted in blue line. The grey shaded area highlights the water vapor activity inactive month with low RMSEs and the red shaded area represents the water vapor activity active month with high RMSEs

as illustrated in Table 3 and Fig. 8, the average IWV value is approximately 20 mm, which aligns with the scales typically observed during the inactive period. To further support this observation, we analyzed the rainfall data for DOY 208 from a daily gridded observational dataset for precipitation, temperature and sea level pressure in Europe (E-OBS)[6] and found that the majority of the region recorded 0 mm of precipitation. This suggests that the overall water vapor activity on that day was low. The relative stability of the water vapor conditions enabled the model to effectively learn the mapping patterns from input data to output data, resulting in superior performance on this day compared to other epochs within the active period. Overall, all correlation coefficients in Table 3 exceed 0.80, indicating consistent trend patterns. Notably, the MBEs are nearly zero, while the RMSEs vary across different periods. For instance, during inactive periods, the RMSEs are smaller than 1 mm, whereas during active periods, the RMSEs range up to 1.5 mm. Moreover, the valid rate values consistently remain high, exceeding 95%.

$$ValidRate(VR) = \frac{N_{all} - N_{outlier}}{N_{all}} * 100\% \quad (31)$$

Additionally, we present the VR and RMSE of the LSSVM model for each experimental epoch throughout the year. In Fig. 10, the blue dots represent the model's VR at each epoch, and the black solid line is the fitted trend line. It is evident that the VRs throughout the year are consistently above 80%, with an average effectiveness rate of 94.4%, demonstrating the model's robustness. A comparison across different times highlights July 2022, marked by the gray area in the figure, where the VR is consistently lower than in other months. Due to heavy rains in July, the region experienced frequent and complex water vapor activities. Consequently, at certain special stations, the water vapor content reached its maximum value, leading to differences when estimating these stations using data from others. Then, the number of outliers increased. Nonetheless, the overall performance and applicability of the model remain strong, as demonstrated in Fig. 11. The red area represents the water vapor active period from May to October (late spring, summer, and early autumn), during which the RMSE values are higher compared to the grey area, which represents the water vapor inactive period (late autumn, winter, and early spring). The average RMSE throughout the year stands at 0.95mm and show this predictive model keep the state-of-the-art level.

Notably, the prediction model maintains its performance even during heavy rainfall events. For instance, on day 201 of 2022, the precipitation data from E-OBS indicates that heavy rainfall occurred in the area. The precipitation data, shown in Fig. 12, demonstrate a multi-center rainfall pattern. Despite these challenging conditions, the prediction results for the GNSS stations within the affected area remained accurate, and there are no outlier stations within the rainfall area. The four outlier stations in the model are presented

[6] Data can be downloaded from https://cds.climate.copernicus.eu/cdsapp#!/dataset/insitu-gridded-observations-europe?tab=overview

in Fig. 12 and Table 4. They reveal that these stations are located at the boundaries of the input training data, including the longitude and altitude, resulting in larger prediction deviations for these stations. This issue arises from the sparse training data at the boundaries. Increasing the amount of training data to achieve a more uniform distribution across the entire range would enhance the model's performance and reduce prediction errors at the edges. However, this evidence still supports the conclusion that the model can sustain its reliability and accuracy even under extreme weather conditions.

Assessment of the LSSVM prediction model with ERA5 data

This study used ERA5 analysis products to compute ERA5 IWV values at each GNSS station location and altitude. Subsequently, the ERA5 IWV dataset is compared with both the predicted IWV from our proposed model and GNSS IWV data. This comparative analysis highlights the superiority of our results of predictive model. Specifically, compared to ERA5 IWV, the IWV generated by our predictive model demonstrates closer proximity to reference GNSS IWV. The corresponding results are depicted in Fig. 13, where the blue line represents ERA5 IWV, the red line represents the model predicted IWV, and the black line represents reference IWV data derived from GNSS observations. In both active and inactive periods, the accuracy of ERA5 IWV is lower than that of the predictive model proposed in this study. The RMSEs are calculated between ERA5 IWV and reference IWV (RMSE_ERA5_REAL), as well as between predicted IWV and reference IWV (RMSE_PREDICTION_REAL) at each epoch. These results are illustrated in Figs. 14 and 15. The average RMSE_ERA5_REAL is 2.41mm, while the average RMSE_PREDICTION_REAL is 0.95 mm. This indicates a considerable improvement and demonstrate that the predictive model generally exhibits a superior estimation process compared to the ERA5 interpolation method. The proposed predictive model achieves a 61% improvement in RMSE compared to the traditional ERA5 calculation model.

Especially during the period from DOY 160 to DOY 250 (July, August, and September), as illustrated in Fig. 14, the

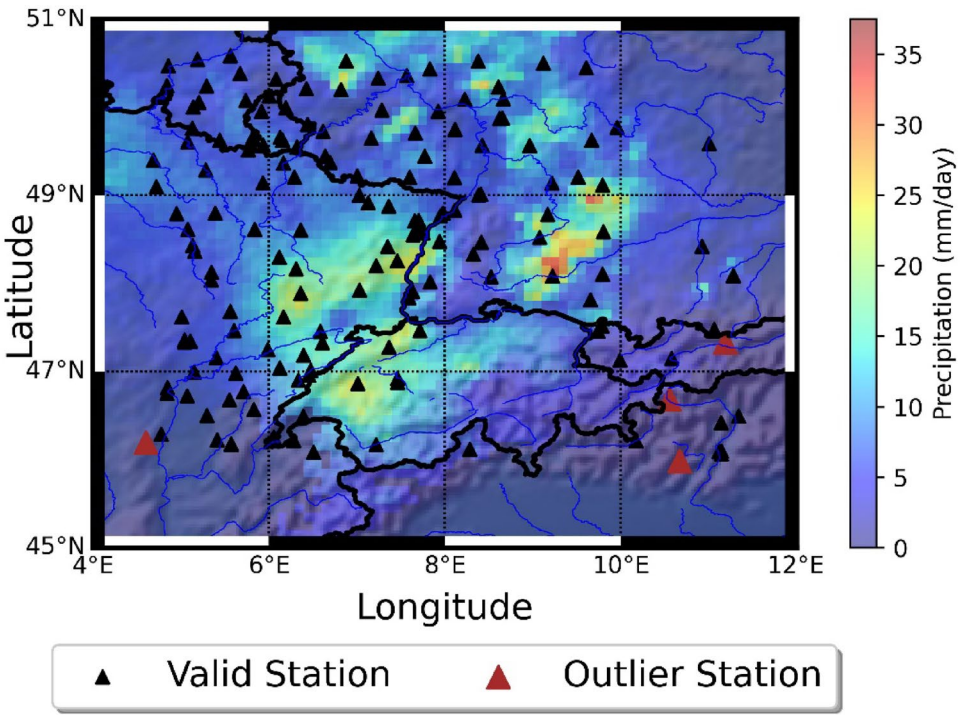

Fig. 12 The predictive results on day 201 of 2022 revealed 4 stations as outliers, with their details presented in Table 4. Additionally, the precipitation conditions for this day are also provided

Table 4 The 4 outlier stations on day 201 of 2022, IWV difference is the difference between the Predicted IWV and Reference IWV

GNSS ID	Longitude (in °)	Latitude (in °)	Altitude (in m)	Mean Temperature (in k)	IWV Difference (in mm)
AVEN	4.60	46.19	740.8	283.1	5.94
MABZ	10.55	46.68	1092	281.8	5.68
RONC	10.66	45.98	885.4	281.5	4.24
SEEF	11.18	47.32	1245	283	4.31

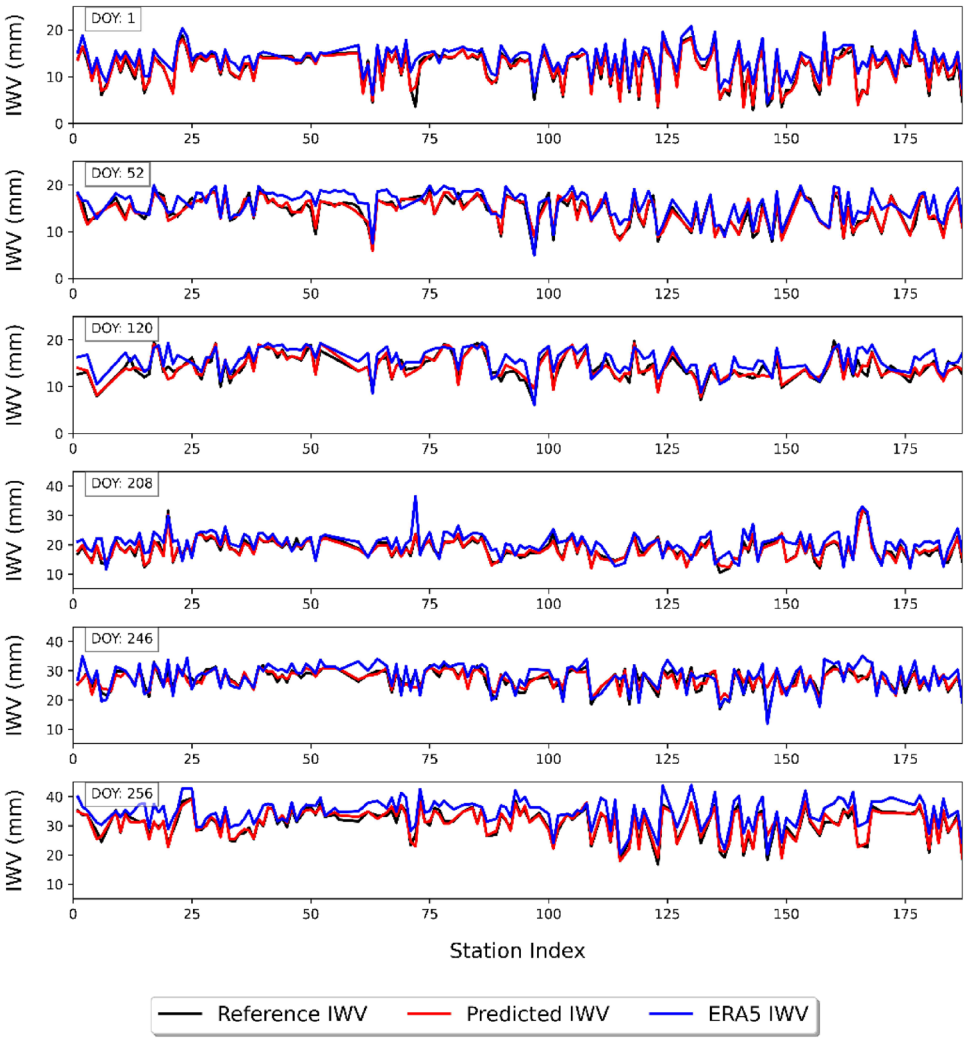

Fig. 13 The series of ERA5 IWV, predicted IWV, and real reference IWV on Day 1, Day 52, and Day 120, representing winter when water vapor activity is typically inactive with IWV range from 0 to 20 mm. Conversely, on Days 208, 246, and 256, corresponding to summer when water vapor activity is active with IWV range from 5 to 45 mm

accuracy of ERA5 in estimating IWV is noticeably lower compared to the proposed model. This difference can be attributed to several factors. A key reason is the increased atmospheric variability during the summer months, which introduces significant complexity to the atmospheric system. This period is characterized by highly dynamic weather patterns, including frequent convective processes and localized precipitation events, which ERA5 often struggles to model accurately. Convective storms and rapid fluctuations in atmospheric moisture can result in substantial discrepancies between ERA5 reanalysis estimates and actual IWV values (Parracho et al. 2018). Furthermore, the observational data used in ERA5 reanalysis, such as satellite and ground-based weather station data, may suffer from limitations in spatial and temporal resolution, exacerbating its inaccuracy under such rapidly changing conditions. In contrast, the proposed LSSVM model shows marked improvement in performance, particularly during the summer months, due to

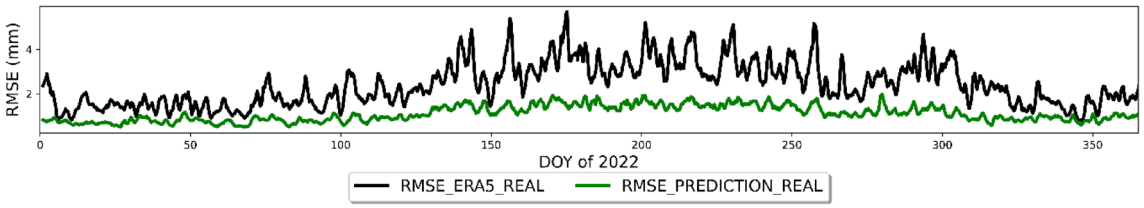

Fig. 14 The comparison of two RMSEs: the black line is the RMSE between ERA5 IWV and real reference IWV, the green line is the RMSE between predicted IWV and real reference IWV

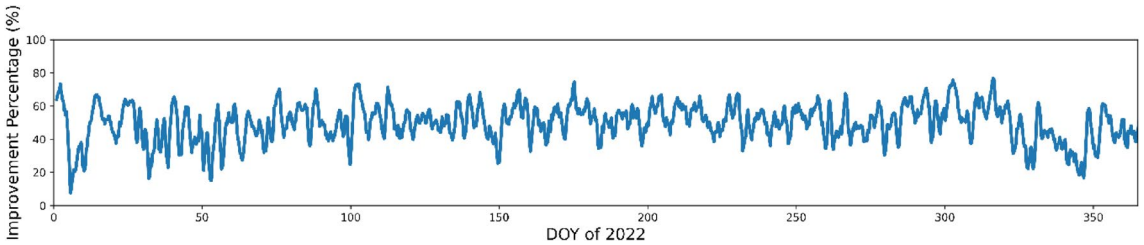

Fig. 15 The percentage of improvement between two kinds of RMSE: the RMSE between ERA5 IWV and real reference IWV, the RMSE between predicted IWV and real reference IWV

its ability to effectively capture the nonlinear and complex relationships present in the data. One of the major strengths of the LSSVM model lies in its use of high-precision GNSS IWV data, which provides a much finer temporal resolution than ERA5. By integrating real-time GNSS IWV measurements into the training dataset, the LSSVM model is better equipped to learn and adapt to the dynamic atmospheric conditions typical of summer, leading to more accurate IWV estimates compared to ERA5.

Overall, the adaptability of the LSSVM model enables it to more accurately estimate IWV even under challenging conditions such as high moisture variability, convective activity, and localized precipitation. This advantage becomes particularly evident during the summer when the atmosphere is more unstable, and the limitations of ERA5 are most pronounced. The LSSVM model's capability to adapt to these conditions allows for a more precise representation of the atmospheric moisture content, significantly reducing the discrepancies between predicted and actual IWV values during this period of high variability.

Assessment of the LSSVM prediction model with radiosonde data

This study computed radiosonde IWV as a reference and calculated the predicted IWV at the positions and altitudes of the radiosonde. Figure 16 present the results of their comparison, where the red line represents the predictions of proposed model and the black line represents the radiosonde IWV. The results for a total of 9 stations are showcased. All cases demonstrate highly similar correlations exceeding 0.95, except for one case nearing 0.90. Table 5 presents the RMSE and correlation of each radiosonde station and the mean RMSE is 2.49 mm. This underscores the predictive capability of the model proposed in this study. As illustrated in Fig. 16, station GMM00010962 exhibits a positive deviation, primarily due to its significantly limited number of data samples compared to other radiosonde stations. This limitation arises from stringent quality control and preprocessing measures that resulted in the exclusion of data points with incomplete meteorological information or fewer than five pressure layers. Consequently, station GMM00010962 yielded only 117 valid data records throughout Year 2022, while other stations typically maintained around 700 records. This restricted dataset leads to a sparse and fragmented time series plot for this station, for instance, only 18 data records were available between DOY 150 and DOY 250, and the connection of scatter points across the time scale visually exaggerates the apparent deviations. Additionally, the radiosonde data from this station generally exhibited poor quality, characterized by a limited number of pressure layers, most often only seven or eight, resulting in an underestimation of the IWV. Nonetheless, the overall trend from station GMM00010962 is consistent with the model outputs, thereby supporting the model's validity. The observed positive deviation can be primarily attributed to the sparse dataset and the inferior quality of radiosonde measurements at this specific station. Moreover, the radiosonde data serves as an external reference, providing high-quality observed IWV that has not been used in the training of the predictive model. This independent dataset helps confirm the performance and reliability of the model.

Sensitivity analysis of LSSVM model to the regularization parameter C

The C is the regularization parameter and plays a key role in modulating the complexity of the model, where higher values of C lead to increased model complexity potentially culminating in overfitting, whereas lower values of C promote model simplicity, albeit at the risk of underfitting. Additionally, the influence of C on model performance is contingent upon the quantity of training samples. Generally, as the sample size expands, smaller values of C tend to be more appropriate, while larger values of C may exacerbate overfitting. In our investigation, we tailored the value of C to different training datasets to ensure the robustness of the corresponding LSSVM models. Here, we present the number of training stations in Fig. 17 and the optimal C values for various training models obtained through our parameter optimizer, as depicted in Fig. 18. Consequently, by contrasting experimental outcomes

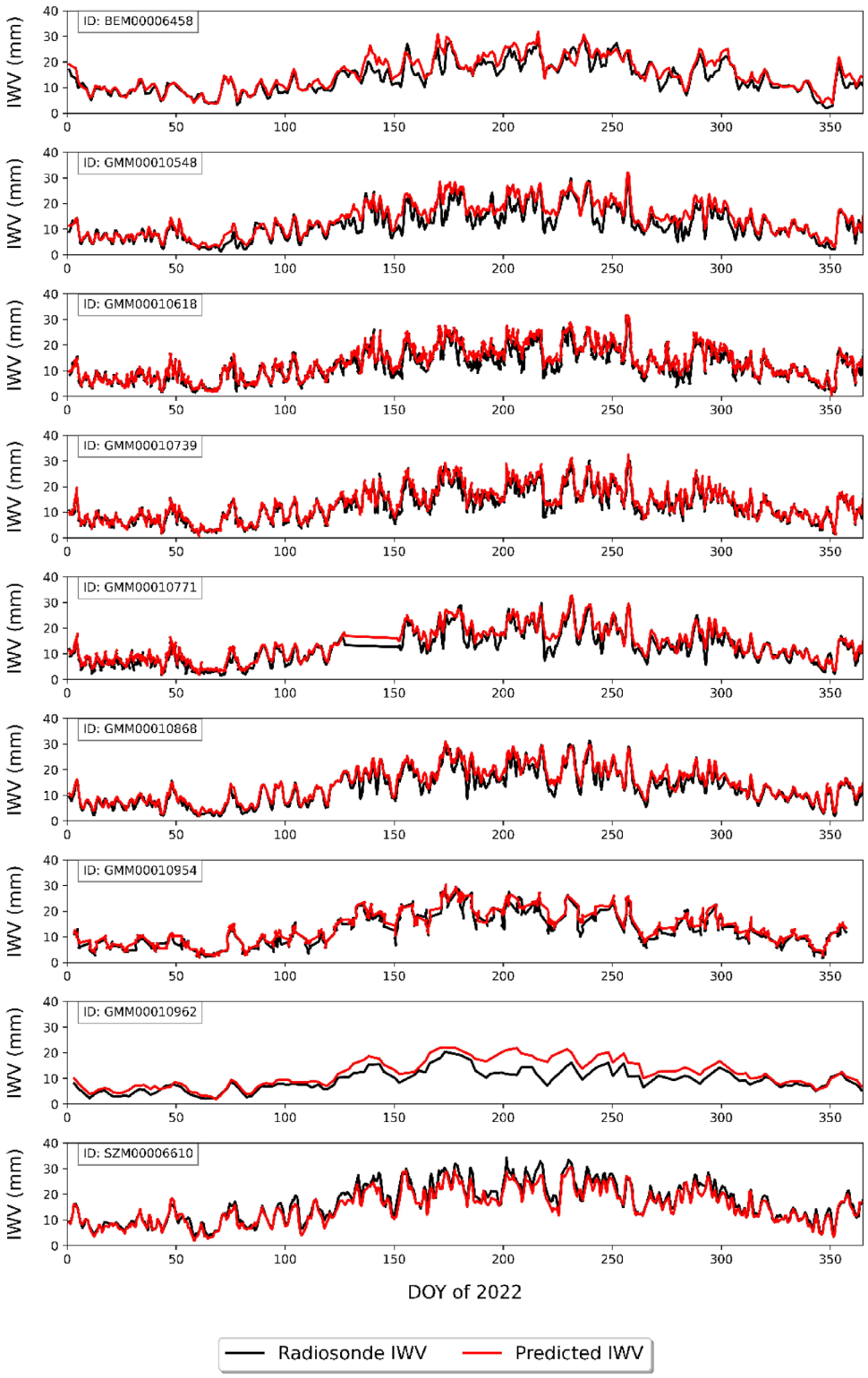

Fig. 16 The comparison between predicted IWV and radiosonde IWV and the text in figure is the radiosonde ID

obtained with non-optimal model (predefining constant C), we illuminated the sensitivity of model estimation results to the selection of C. The adoption of an appropriate C value holds the potential to substantially reduce RMSE in Figs. 19 and 20. After comparison, it is evident that the optimal parameter C can reduce the RMSE, underscoring its significance in enhancing IWV prediction accuracy.

Table 5 The RMSE and correlation of validation with radiosonde data

Radiosonde ID	RMSE (in mm)	R	Radiosonde ID	RMSE (in mm)	R
BEM00006458	3.09	0.98	GMM00010868	2.32	0.96
GMM00010548	2.43	0.95	GMM00010954	2.06	0.96
GMM00010618	2.03	0.97	GMM00010962	2.75	0.99
GMM00010739	2.84	0.99	SZM00006610	2.58	0.90
GMM00010771	2.34	0.96	Mean	2.49	0.96

Performance of the LSSVM model during tornado events and heavy rains

The performance of the proposed LSSVM model was evaluated under extreme weather conditions characterized by significant atmospheric disturbances, specifically focusing on tornado events and heavy rainfall within the Upper Rhine Graben region. Tornadoes, known for their rapid and extreme variations in water vapor content, present a critical challenge for assessing the robustness and adaptability of predictive models. We identified three tornado cases during the study period and analyzed the model's capability to accurately estimate IWV in the face of these rapid atmospheric fluctuations. For each event, the LSSVM model's IWV predictions were compared with ground truth data from GNSS observations. Additionally, the model's performance was examined during periods of heavy rainfall, which are marked by sustained and intense precipitation, resulting in sharp increases in atmospheric moisture content. Three heavy rainfall events were selected for analysis to further assess the model's ability to handle rapid water vapor variations. All cases are listed in Table 6.

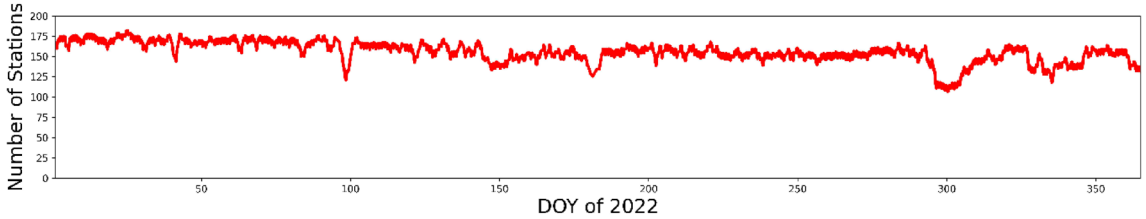

Fig. 17 The number of training stations in different training models

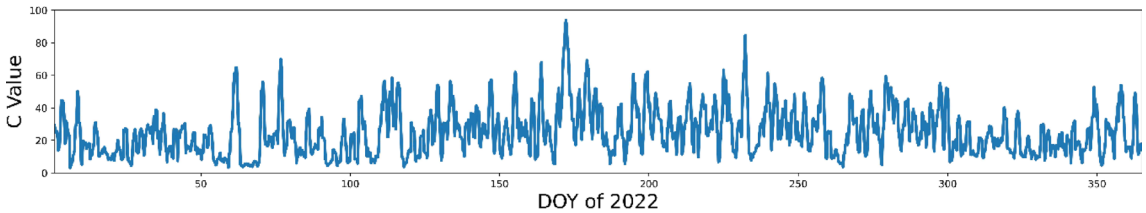

Fig. 18 The optimal C values in different training models

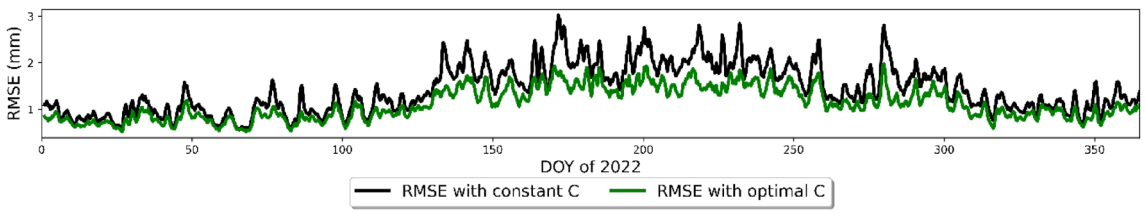

Fig. 19 The comparison of two RMSEs: the black line is the RMSE of model with optimal C values, the green line is the RMSE of model with non-optimal C values

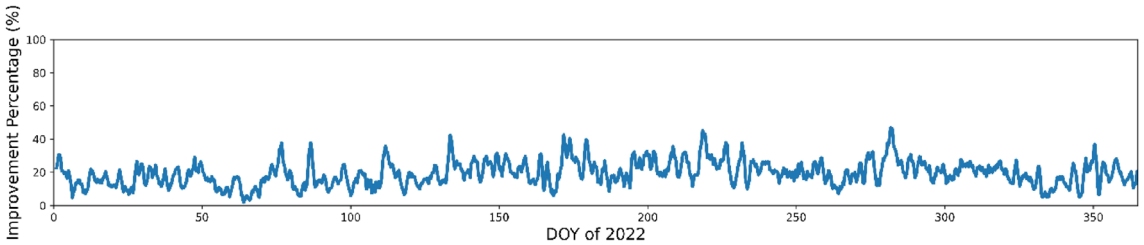

Fig. 20 The percentage of improvement between two kinds of RMSE: the RMSE of model with optimal C values, the RMSE of model with non-optimal C values

Table 6 The validation cases for three tornado events and three heavy rains with event information sourced from the European Severe Weather Database (ESWD)

Event Type	City	Location	DOY of 2022	Time
Tornado	Böhmenkirch, Germany	48.68 N, 9.93 E	99	17:17 UTC
Tornado	Landersheim, France	48.69 N, 7.50 E	125	14:00 UTC
Tornado	Urexweiler, Germany	49.43 N, 7.08 E	322	13:45 UTC
Heavy Rain	Primstal, Germany	49.53 N, 6.98 E	247	13:30 UTC
Heavy Rain	Virton, Belgium	49.56 N, 5.52 E	258	16:54 UTC
Heavy Rain	Châtel, France	46.27 N, 6.84 E	298	14:05 UTC

Data can be downloaded from https://eswd.eu/cgi-bin/eswd.cgi

In Fig. 21, we observe that across the three tornado cases, the LSSVM estimation model demonstrated strong accuracy compared to GNSS IWV values. The results show that the model performed admirably, with RMSE values of 0.65 mm, 0.82 mm, and 0.71 mm, respectively, successfully capturing the rapid variations in IWV associated with these phenomena. The model effectively adapted to the localized impacts of tornadoes, displaying minimal deviation from the actual measurements. Similarly, during heavy rainfall events, the model maintained a high level of accuracy, with RMSE values of 1.24 mm, 1.67 mm, and 1.03 mm, highlighting its consistency in predicting IWV levels under challenging weather conditions. However, the accuracy of the LSSVM model in heavy rain events is slightly lower than in tornadoes, likely due to the sharp increase in water vapor during rainfall, while tornadoes typically result in rapid decreases in atmospheric moisture. Nevertheless, the model's robust performance in handling rapid water vapor variations across both types of extreme weather demonstrates its significant potential for broader application in various meteorological contexts.

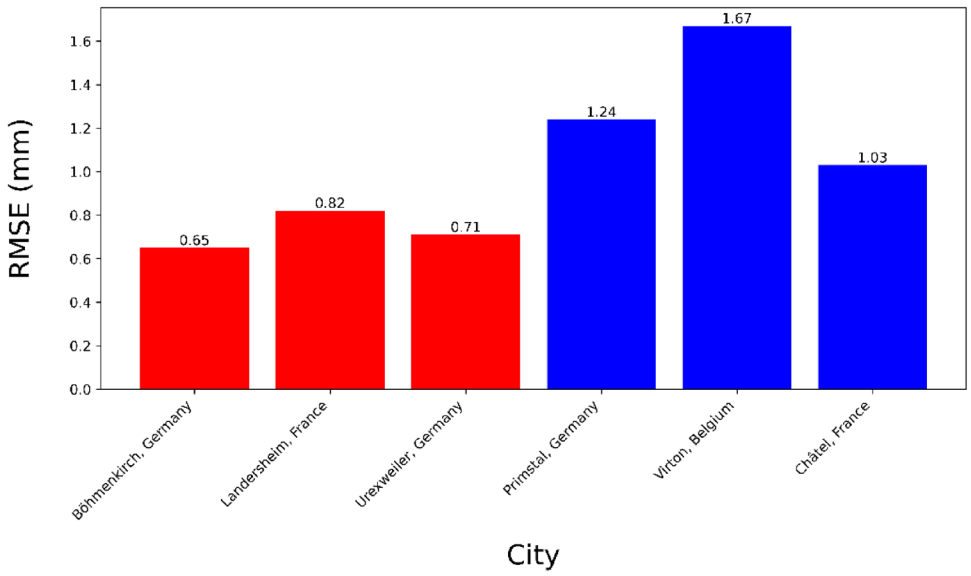

Fig. 21 The validation results cover six extreme weather events in 2022: the red bars represent tornado cases, while the blue bars correspond to heavy rain events

IWV variation maps across the region

Following comprehensive validation and evaluation, the proposed LSSVM predictive model was utilized in this study to forecast IWV values across the entire study area. A key strength of the model is its adaptability to varying terrains, which significantly influences its overall performance. The experimental region, the Upper Rhine Graben in Central Europe, encompasses a geographically diverse landscape, including plains, mountainous areas, and low-lying regions. As detailed in Section "Case Study", the elevation within this region spans from sea level to over 4000 m, offering a complex and varied topographical environment for model testing. Despite these challenges, the model consistently demonstrated high accuracy across different seasons, further underscoring its robustness in capturing IWV variability in diverse terrains. Subsequently, these forecasts were visualized to generate comprehensive maps, showcasing the model's efficacy for short-term climate analysis at hourly intervals, as delineated in Fig. 22. Throughout this process, the model utilized IWV data from various GNSS stations on an hourly basis to generate predictive IWV distribution maps based on a DEM. This systematic approach underscored the model's effectiveness in short-term studies. Specifically, we presented IWV maps for four distinct time intervals within a single day, elucidating the intricate variations in IWV across the region throughout the day. These findings show the model's capability to accurately capture short-term IWV trends within the region. Moreover, our model demonstrated stability and effectiveness in the long-term monitoring of water vapor variation, as illustrated in Fig. 23. In contrast to short-term input data, long-term investigations can leverage daily or monthly averaged IWV values for prediction. Here, we present the map of daily averaged IWV for the region,

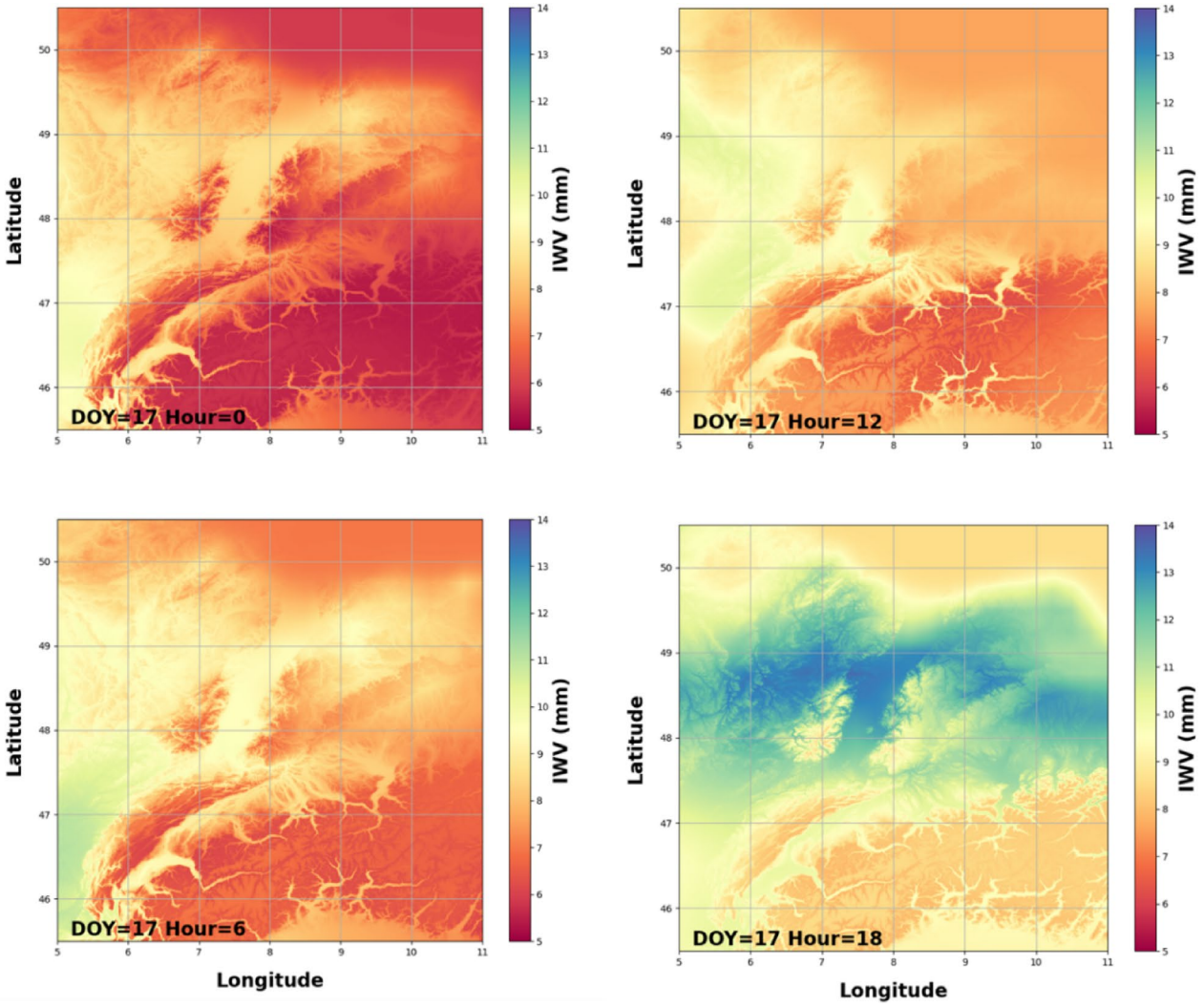

Fig. 22 The hourly IWV distribution maps across the region at hours 0, 6, 12, and 18 on Day of Year (DOY) 17 show the following uncertainties: 0.58 mm, 0.65 mm, 0.62 mm, and 0.72 mm, respectively

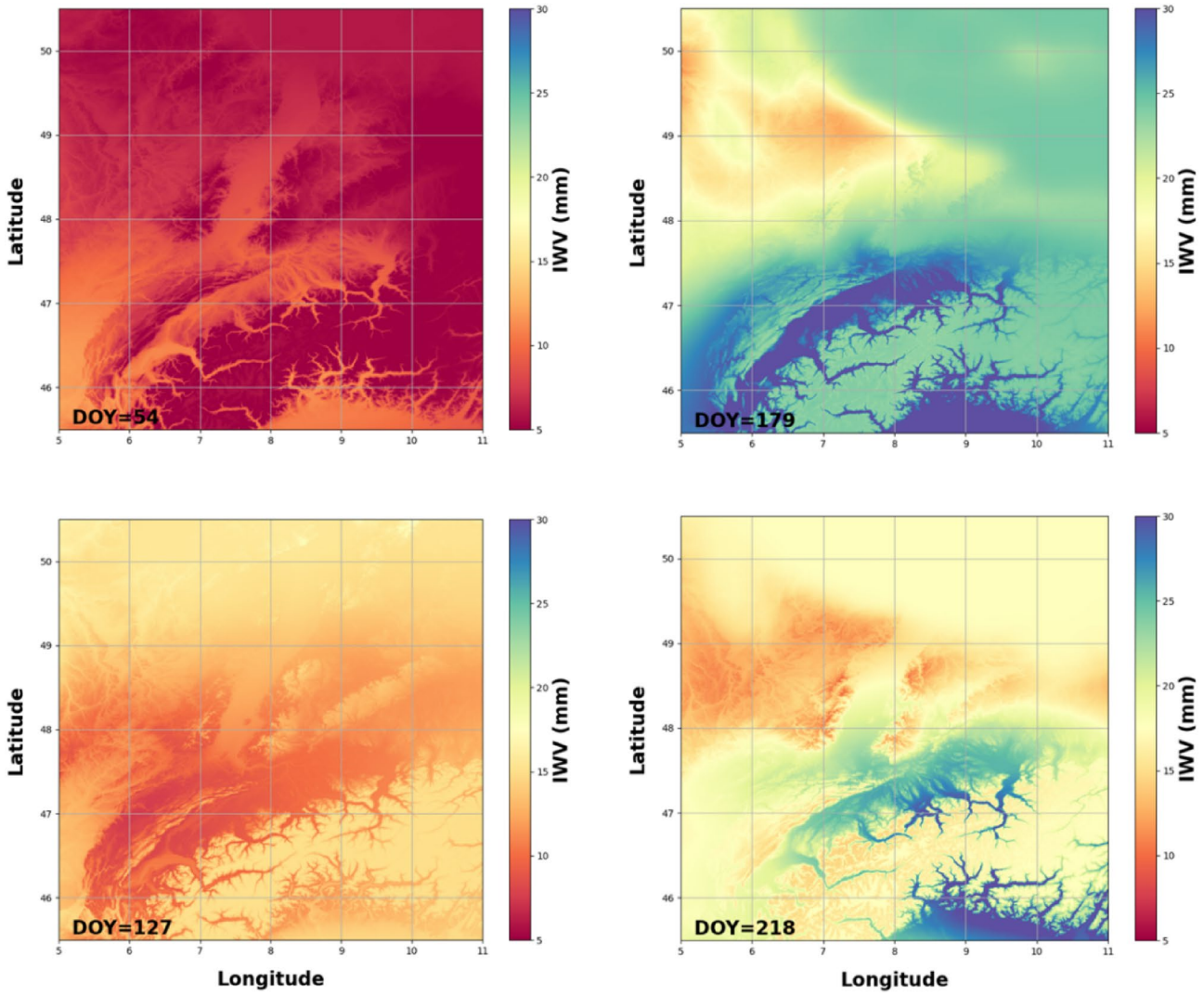

Fig. 23 The daily IWV distribution maps across the region of Day of Year (DOY) 54, 127, 179 and 218 show the following uncertainties: 0.66 mm, 0.72 mm, 1.26 mm, and 1.36 mm, respectively

which can be beneficial in exploring annual variations in water vapor content and climate changes within the region.

Conclusion

In conclusion, this study presents an improvement in IWV estimation through introduction of a regional IWV predictive model based on Machine Learning. The proposed model, an advanced LSSVM predictive model, autonomously determines optimal parameters, enhancing predictive accuracy. Leveraging IWV data from diverse GNSS stations, the model offers accurate IWV estimation at any location within the region, addressing the spatial coverage limitation associated with traditional methods. Remarkably, the model achieves an overall average RMSE of 0.95 mm, indicating its robust predictive capability. Furthermore, the performance of model demonstrates adaptability across different seasons and terrains, highlighting its versatility in diverse environmental conditions. By evaluating the reliability of the conventional ERA5 IWV calculation method against the proposed model, a notable 61% improvement rate is observed. These findings underscore the practical utility of the developed model in environmental research for the Upper Rhine Graben Region. Overall, this study contributes to advancing IWV estimation techniques, offering a reliable and accurate predictive model that can be applied effectively in regions lacking IWV measurement infrastructure. Our future work will focus on developing a real-time predictive model by integrating real-time GNSS data with real-time meteorological datasets. This approach aims to enhance the timeliness and accuracy of IWV real time predictions.

Acknowledgements The authors would like to thank the NGL (http://geodesy.unr.edu) for providing experiment data. The first author was financed by China Scholarship Council (CSC). This research was partly financed by the Institute of Geodesy (GIK) of Karlsruhe Institute for Technology. We sincerely thank the reviewers for their valuable feedback and suggestions.

Author contributions Lingke Wang wrote the manuscript text and prepared figures. All authors reviewed the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

Akar AU, Inal C (2023) Prediction of Zenith tropospheric delay in GNSS observations using support vector regression. Adv Space Res 71(11):4659–4680

Askne J, Nordius H (1987) Estimation of tropospheric delay for microwaves from surface weather data. Radio Sci 22(03):379–386

Baba Shaeb Kannemadugu H, Ranganathan K, Gharai B, Seshasai MVR (2022) GNSS-GPS derived integrated water vapor and performance assessment of ERA-5 data over India. J Atmos Solar-Terr Phys 227:105807

Bevis M, Businger S, Herring TA, Rocken C, Anthes RA, Ware RH (1992) GPS meteorology: remote sensing of atmospheric water vapor using the global positioning system. J Geophys Res: Atmo 97(D14):15787–15801

Bevis M, Businger S, Chiswell S, Herring TA, Anthes RA, Rocken C, Ware RH (1994) GPS meteorology: mapping zenith wet delays onto precipitable water. J Appl Meteorol 33:379–386

Blewitt G, Hammond W (2018) Harnessing the GPS data explosion for interdisciplinary science. Eos. https://doi.org/10.1029/2018EO104623

Boehm J, Niell A, Tregoning P, Schuh H (2006a) Global mapping function (GMF): a new empirical mapping function based on numerical weather model data. Geophys Res Lett. https://doi.org/10.1029/2005GL025546

Böhm J, Werl B, Schuh H (2006b) Troposphere mapping functions for GPS and VLBI from ECMWF operational analysis data. J Geophys Res 111(B2):B02406

Chen G, Herring T (1997) Effects of atmospheric azimuthal asymmetry on the analysis of space geodetic data. J Geophys Res: Solid Earth 102(B9):20489–20502

Durre I, Xungang Y, Vose RS, Applequist S, & Arnfield J (2016) Integrated global radiosonde archive (IGRA) Version 2. NOAA national centers for environmental information, 10, V5X63X0Q.

Huang L, Wang X, Xiong S, Li J, Liu L, Mo Z, Fu B, He H (2022) High-precision GNSS PWV retrieval using dense GNSS sites and in-situ meteorological observations for the evaluation of MERRA-2 and ERA5 reanalysis products over China. Atmos Res 276:106247

Izanlou S, Haji-Aghajany S, Amerian Y (2024) Enhanced troposphere tomography: integration of GNSS and remote sensing data with optimal vertical constraints. IEEE J Sel Top Appl Earth Obs Remote Sens 17:3701–3714. https://doi.org/10.1109/JSTARS.2024.3354884

Jade S, and Vijayan MSM (2008) GPS-based atmospheric precipitable water vapor estimation using meteorological parameters interpolated from NCEP global reanalysis data. Journal of Geophysical Research: Atmospheres, 113(D3). https://doi.org/10.1029/2007JD008758

Jiang C, Xu T, Wang S, Nie W, Sun Z (2020) Evaluation of zenith tropospheric delay derived from ERA5 data over China using GNSS observations. Remote Sens 12(4):663

Jiang C, Gao X, Wang S, Zhu H, Xu A, An Q, Zhu M, Liu G (2023) Comparison of ZTD derived from CARRA, ERA5 and ERA5-Land over the Greenland based on GNSS. Adv Space Res 72(11):4692–4706

Kestin J, Sengers JV, Kamgar-Parsi B, Sengers JMHL (1984) Thermophysical properties of fluid H_2O. J Phys Chem Ref Data 13:175–183

Li S, Xu T, Jiang N, Yang H, Wang S, Zhang Z (2021) Regional zenith tropospheric delay modeling based on least squares support vector machine using GNSS and ERA5 data. Remote Sens 13(5):1004

Lu C, Zheng Y, Wu Z, Zhang Y, Wang Q, Wang Z, Zhong Y (2023) TropNet: a deep spatiotemporal neural network for tropospheric delay modeling and forecasting. J Geod 97(4):34

Mehrkanoon S, Falck T, Suykens JA (2012) Parameter estimation for time varying dynamical systems using least squares support vector machines. IFAC Proc Vol 45(16):1300–1305

Migliaccio M, Gambardella A (2005) Microwave radiometer spatial resolution enhancement. IEEE Trans Geosci Remote Sens 43(5):1159–1169

Namaoui H, Kahlouche S, Belbachir AH (2021) Evaluation of MODIS water vapour products over ALGERIA using radiosonde data. Anuário Do Instituto De Geociências. https://doi.org/10.11137/1982-3908_2021_44_40110

Parracho AC, Bock O, Bastin S (2018) Global IWV trends and variability in atmospheric reanalyses and GPS observations. Atmos Chem Phys 18(22):16213–16237

Peters G, van Balen RT (2007) Tectonic geomorphology of the northern Upper Rhine graben. Ger Glob Planet Change 58(1–4):310–334

Prasad AK, Singh RP (2009) Validation of MODIS Terra, AIRS, NCEP/DOE AMIP-II reanalysis-2, and AERONET Sun photometer derived integrated precipitable water vapor using ground-based GPS receivers over India. J Geophys Res: Atmos. https://doi.org/10.1029/2008JD011230

Razin MRG, Voosoghi B (2022) Modeling of precipitable water vapor from GPS observations using machine learning and tomography methods. Adv Space Res 69(7):2671–2681

Rocken C, Ware R, Van Hove T, Solheim F, Alber C, Johnson J, Bevis M, Businger S (1993) Sensing atmospheric water vapor with the global positioning system. Geophys Res Lett 20(23):2631–2634

Saastamoinen J (1972) Atmospheric correction for the troposphere and stratosphere in radio ranging satellites. Use Artif Satell Geod 15:247–251

Seidel DJ, Sun B, Pettey M, Reale A (2011) Global radiosonde balloon drift statistics. J Geophys Res: Atmos. https://doi.org/10.1029/2010JD014891

Ssenyunzi RC, Oruru B, D'ujanga FM, Realini E, Barindelli S, Tagliaferro G, Engeln AV, van de Giesen N (2020) Performance of ERA5 data in retrieving precipitable water vapour over east African tropical region. Adv Space Res 65(8):1877–1893

Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9:293–300

Thierion C, Longuevergne L, Habets F, Ledoux E, Ackerer P, Majdalani S, Leblois E, Lecluse S, Martin E, Queguiner S, Viennot P (2012) Assessing the water balance of the Upper Rhine Graben hydrosystem. J Hydrol 424:68–83

Timar-Geng Z, Fügenschuh B, Wetzel A, Dresmann H (2006) Low-temperature thermochronology of the flanks of the southern Upper Rhine Graben. Int J Earth Sci 95:685–702

Vapnik V, Golowich S, Smola A (1996) Support vector method for function approximation, regression estimation and signal processing. Advances in Neural Information Processing Systems, pp. 281–287.

Wang H, Hu D (2005). Comparison of SVM and LS-SVM for regression. In: 2005 International conference on neural networks and brain Vol. 1, pp. 279–283. IEEE.

Wang Z, Chen S (2007) New least squares support vector machines based on matrix patterns. Neural Process Lett 26:41–56

Wilgan K, Hadas T, Hordyniec P, Bosy J (2017) Real-time precise point positioning augmented with high-resolution numerical weather prediction model. GPS Solut 21:1341–1353

Yuan P, Blewitt G, Kreemer C, Hammond WC, Argus D, Yin X, Malderen RV, Mayer M, Jiang W, Awange J, Kutterer H (2022) An enhanced integrated water vapour dataset from more than 10,000 global ground-based GPS stations in 2020. Earth Syst Sci Data Discuss 2022:1–33

Zhang M, Wang M, Guo H, Hu J, Xiong J (2023) Tropospheric delay model based on VMF and ERA5 reanalysis data. Appl Sci 13(9):5789

Zhang M, Li L, Zhang K, Wu S, Sun P, Zhao D, Rohm W (2024) Development of an adaptive 4-D water vapour density model for the vertical constraints in GNSS tropospheric tomography. GPS Solut 28(4):159

Lingke Wang is currently a Ph.D. candidate at Karlsruhe Institute of Technology (KIT), Germany. He received his master's degree from University of Stuttgart, Germany in 2021 and received his bachelor's degree from Wuhan University, China in 2019. His research interests include GNSS data processing, tropospheric water vapor model and multipath effect mitigation model.

Hansjörg Kutterer is University Professor for Geodetic Earth System Science at the Karlsruhe Institute of Technology, Germany, since 2018. He received his Dipl.-Ing., Dr.-Ing. and Dr.-Ing. habil. degrees at the University of Karlsruhe in 1990, 1993 and 2001, respectively. From 2004 to 2011 he was University Professor at Leibniz University Hannover and from 2011 to 2018 he was Director General of the Federal Agency of Cartography and Geodesy in Frankfurt/Main, Germany. He has been holding various high-level positions in national and international administrative and scientific committees. At present, e.g., he is Assistant Editor-in-Chief of the Journal of Geodesy. His research interests refer to satellite-based geodetic observation techniques and systems such as GNSS and InSAR, to geodetic modelling and data analysis and their relation to Earth system research.