https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

Open Access Earth System Science Data Discussions

A high-quality gap-filled daily ETo dataset for China during 1951-2021 from synoptic stations using machine learning models

NingShan Zhou[1], LiFeng Wu[1,2,3*], QiLiang Yang [1,2,3*], Jianhua Dong[4*], Ling Yang[5], Yue Li[1,2,3]

[1]Faculty of Modern Agricultural Engineering, Kunming University of Science and Technology, Kunming, 650500, Yunnan, China
[2]Yunnan Provincial Field Scientific Observation and Research Station on Water-Soil-Crop System in Seasonal Arid Region, Kunming University of Science and Technology, Kunming, 650500, Yunnan, China
[3]Yunnan Provincial Key-Laboratory of High-Efficiency Water Use and-Green Production of Characteristic Crops in Universities, Kunming University of Science and Technology, Kunming, 650500, Yunnan, China
[4]State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China
[5]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, Yunnan, China.

Correspondence to: LiFeng Wu(china.sw@163.com), Qiliang Yang (yangqilianglovena@163.com), Jianhua Dong (djh0530dyz@126.com)

Abstract. The reference evapotranspiration (ETo) is essential for water-consuming in agriculture and land-water cycle research. The synoptic data from meteorological stations can provide reliable ground data for ETo estimation with the FAO-56 Penman-Monteith equation. However, the five primary variables this equation needs, including maximum temperature (Tmax), minimum temperature (Tmin), sunshine duration (SSD), wind speed (Wind), and relative humidity (RH), often experience severe data loss due to force majeure events in synoptic data. The data loss would directly introduce severe data gaps to the complex records for ETo. Machine learning algorithms can fill various data gaps with low error rates, however, to achieve high data quality, the algorithms must be selected properly to deal with the distinct types of data loss and train independently. Here, based on the data characters, we investigated and classified data gaps from the synoptic dataset into 2 major types: the common, minor data loss gaps including Tmax loss/Tmin loss/SSD loss/Wind loss/RH loss/Wind and SSD loss/Wind and RH loss, and the other 19 types of data loss which is more severe in information loss but barely occurred. Our results show that the XGBoost model achieved the best accuracy in all 3 machine learning models with high statistic levels. For the other 19 types of data gaps, the LSTM models were trained separately for each site and achieved average R^2, RMSE, and nRMSE at 0.9, 0.5 mm d^{-1}, and 38% for the total 2419 stations. Thus, we propose a high-quality, gap-filled daily ETo dataset during 1951-2021 for China with the proportion of large errors (the data with daily ETo errors more than 1.5 mm d^{-1}) below 0.2%. Our results also reveal that the entanglement degree between synoptic variables varies a lot from region to region in China. Although most research indicates that wind speed is not very important for ETo estimation with machine learning models, our findings reveal that wind speed played a more significant role in ETo estimation in most areas of China during the years before the 21st century. Still, the impact of wind speed on ETo has also been alleviated in recent years. This ETo dataset for China is available online at https://doi.org/10.5281/zenodo.11496932 (Zhou et al., 2024).

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

1 Introduction

Evapotranspiration (ET) is the fundamental process of water loss in agricultural fields (Tanner, 1967), and as a crucial component of the ecosystem health, hydrological cycle, evaporation influence deeply in water resources preservation, crop yield irrigation, water management practices (Wanniarachchi & Sarukkalige, 2022). Understanding the impact of various factors on ET is crucial, as ETo, or reference evapotranspiration, serves as a standardized benchmark for comparing and predicting ET across different regions and conditions. (Chen & Liu, 2020; Elhaddad & Garcia, 2008; Gowda et al., 2008). ETo provides a standardized measure that can be adjusted for local conditions and specific crop water consumption (Shiri, 2017). In the world wild, the FAO-56 PM equation is the benchmark ETo estimation method for crops in various soil conditions (Fan & Thomas, 2013; Li et al., 2012). The computation of the FAO-56 PM equation required detailed environmental data, including solar radiation (Rs), maximum and minimum temperatures (Tmax/Tmin), relative humidity (RH), and wind speed at a 2-meter height (WS) (L. S. Pereira et al., 2015). Based on the different sources of climate data, multiple ETo data products are provided in worldwide.

In recent years, both satellite-based and ground-information-based ETo products have been made available globally by multiple suppliers, offering ETo datasets at varying temporal and spatial scales. Satellite-based ETo products like the Global Land Data Assimilation System (GLDAS), Global Land Evaporation Amsterdam Model (GLEAM), and the Numerical Terradynamic Simulation Group (NTSG) provided daily ETo at spatial resolution ranging from 10km to 0.25° from 1982 (Bai & Liu, 2018). The ground-information-based ETo products can offer higher spatial-temporal resolution than the satellite products, for example, the geographic remote sensing ecological network offered high-resolution ETo products from 1958 to recent years based on the climate data spatial interpolation software Anusplin to predict the actual evapotranspiration data for China at hourly, daily, monthly, and yearly scale at spatial resolution range from 10m to 1km. But both the satellite and ground-information-based ETo products have severe limits of error and uncertainty caused by various data collection and recording methods (Bormann et al., 1996; Ghilain et al., 2011). Changes in data collection methods and equipment updates over a long period would also lead to inevitable data loss and fuzzy (Gavilán & Castillo-Llanque, 2009; Paredes & Pereira, 2019). These inevitable factors have led to significant data gaps in the synoptic dataset, hindering the further computation and application of long-term ETo values, accurate ETo prediction over large regions, and effective agricultural irrigation management in certain areas (Z. Hu et al., 2022; Malik et al., 2022; Roy, 2021).

Although it is hard to fill the synoptic data gaps, the ETo data gaps could be filled by both empirical equations and machine learning models (Gocic et al., 2016). Empirical ETo estimation models are based on simulating the physical evaporation process and energy cycles. The most canonical equations for ETo estimation under limited data are the Hargreaves Equation (Hargreaves & Allen, 2003), Thornthwaite Equation (CHANG, 1959), Blaney-Criddle Equation (Allen & Pruitt, 1986), and the Priestley-Taylor Equation (LHOMME, 1997). The Hargreaves Equation required mainly temperature, and the Thornthwaite Equation calculates ETo based primarily on temperature and day length, while the Blaney-Criddle Equation achieved more accuracy compared to the Thornthwaite Equation in arid regions. The Priestley-Taylor equation required solar

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

radiation, temperature, and VPD to calculate ETo. Although these equations all request fewer environmental parameters than the PM equation, these equations have their limitations in precisions and applicable geographical scope (Kra, 2010; Mallikarjuna et al., 2014; Valiantzas, 2018). In all, the Priestley-Taylor equation could replace the FAO-56 PM Equation only in arid or semi-arid regions where evaporation is often controlled by moisture rather than energy availability (Shiri, 2017). These limitations restricted the usage scenarios of empirical methods.

Compared to the empirical methods, machine learning algorithms could achieve high-quality regression processes under sufficient information and simulate different types of data automatically (Kinaneva et al., 2021). These methods are famous for their robustness and convenience in computation. Thus, they are wildly used to predict ETo from limited environmental data across the world (Mostafa et al., 2023). When the historical climate data are comprehensive enough to encapsulate environmental changes within a specified location, the regression analysis results will be robust and reliable for these regions (Hossein Kazemi et al., 2020; Z. Hu et al., 2022; Mostafa et al., 2023; Santos et al., 2023). Almost all machine learning algorithms can model regression relationships between input and output. This capability allows machine learning models to extract patterns from limited synoptic variable inputs and subsequently infer the relationship between ETo and constrained environmental data. (Granata, 2019). The performance of these machine learning algorithms is significantly influenced by the completeness and quality of the input data (Huang et al., 2020). The data quality and the algorithm's extraction ability to the data features become the two critical points in machine learning approaches (Kim et al., 2022). Data quality is controlled by the data provider, a reliable provider could consume the reliability of synoptic data. As for the algorithms, based on the underlying algorithm principles, the machine learning algorithm used in evaporation regression could be classified to 6 major types: Neural Networks (mainly including Artificial Neural Network short for ANN, Multilayer Perceptron short for MLP, Radial Basis Function short for RBF, Generalized Regression Neural Network short for GRNN, Long Short-Term Memory short for LSTM) (Kisi, 2008), Ensemble Methods (including Random Forests short for RF, Light Gradient Boosting Machine short for Light-GBM, Extreme Gradient Boosting short for XGBoost, M5 Model Tree short for M5Tree) (Salahudin et al., 2023), Fuzzy Systems (Adaptive Neuro-Fuzzy Inference System short for ANFIS) (Ladlani et al., 2014), Genetic Programming (Gene Expression Programming short for GEP, Genetic Programming short for GP) (Guven et al., 2008) , Support Vector Regression short for SVR (Chia et al., 2020) and other learning algorithms like Extreme Learning Machine short for ELM (Gocic et al., 2016), Multivariate Adaptive Regression Splines short for MARS (Kisi, 2016), Cuckoo Search Algorithm short for CSA (Shamshirband et al., 2016). From research for ETo estimation, all the machine learning models based on the 6 types of algorithm principles could achieve high performance, though there is some difference among all algorithms (Granata, 2019). When data fluctuations become more severe than usual, particularly in those areas where perception and high temperature are less relative to the humidity, the GBDT algorithm would perform better because it could handle the outliers and more irregular data (Huang et al., 2020). Also, Cubist presented high performance in modeling daily ETo.

While these algorithms have demonstrated significant potential, their effectiveness is still influenced by various factors including prediction time scales and regional data characteristics. It is crucial to evaluate the performance across different time scales and regional contexts for machine learning models. Most research focuses on ETo prediction periods of 1-3 days, when

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

the prediction period exceeds 3 days, accuracy declines rapidly. For prediction periods longer than 4-7 days, utilizing forecasted weather data to predict ETo is a more suitable approach. This method helps maintain the reliability of ETo predictions over extended periods, leveraging more accurate and updated weather forecasts to improve the overall prediction accuracy. . In the newest research for large climate models, future prediction based on machine learning algorithms could forecast the environmental data for up to 10 days or no longer (Bi et al., 2023). As for the time scale of more than a month and annual prediction, it has been demonstrated that machine learning algorithms can produce results at a monthly scale for 1-3 months and on an annual scale for 1-3 years. (A. R. Pereira & Pruitt, 2004). Another issue in ETo prediction models is that, even with the same data quality and quantity, the performance of a machine learning algorithm also depends on the regional characteristics of the data (Salahudin et al., 2023). As the Priestley-Taylor equation simulates ETo more effectively in arid or semi-arid regions, the quality of the simulation is also influenced by regional data and data characteristics. Lower frequencies of extreme values (primarily maximum values) and outliers facilitate model stability. Consequently, models may exhibit less precise performance in certain seasons within regions characterized by high precipitation and soil moisture deficiency, where data variability is more significant.

To creat a high precision, complex daily ETo dataset for China, we selected a complex meteorological dataset to provide the primary ground information. The data gaps in this meteorological are analysed and classified into 27 types and further classified into 7 major data loss types and other data loss types for the ETo filling. The MARS, SVR, and XGBoost algorithms, which do not account for temporal relationships in the data, are employed to create the machine learning models to fill gaps where one or two daily variables are missing, even when these gaps extend over long periods. For data gaps involving more than three missing indices, the LSTM algorithm is utilized, as it performs better in cases of extensive data loss, except for prolonged continuous data loss. Notably, long-term gaps with more than three missing daily variables occur only in rare situations. All four machine learning models used in gap-filling tasks—MARS, SVR, XGBoost, and LSTM—are reliable according to statistic estimations. By addressing data gaps based on their specific loss types, we established a high-precision ETo dataset suitable for practical applications and ETo data analysis tasks. Additionally, we analysed the importance distribution for ETo prediction under various parameter loss scenarios across 2419 sites and found that the interdependence of different variables varies by region. Among all environmental parameters, Wind is particularly noteworthy; reducing the input data length for Wind results in improved simulation performance regardless of the machine learning algorithm used. This phenomenon may be linked to the decreased independence of the Wind parameter.

where ΔM_0 is Ut rutrum, sapien et vulputate molestie, augue velit consectetur lectus, bibendum porta justo odio lobortis ligula. In in urna nec arcu iaculis accumsan nec et quam. Integer ut orci mollis, varius justo vitae, pellentesque leo. Ut.

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

2 Dataset and methods

2.1 Dataset

2.1.1 Meteorological Dataset

The information source used to calculate the FAO-56 PM equation ETo must provide the following 5 meteorological variables simultaneously: maximum temperature (Tmax), minimum temperature (Tmin), solar radiation (Rs), relative humidity (RH), and wind speed (Wind). Therefore, the meteorological dataset from the National Climatic Centre of the China Meteorological Administration (NCC-CMA) is chosen for the computation process. This dataset provides daily atmospheric data collected from 2,419 meteorological stations across China during 1951 to 2021. The meteorological dataset demonstrates reliability, exhibiting errors within acceptable margins and high precision on a daily scale. This dataset provide eight essential climate data metrics for each ETo record including site ID (ID), longitude (Lon), latitude (Lat), altitude (Alt), year (Year), month (Month), day (Day), Tmax, Tmin, sunshine duration (SSD), relative humidity (RH), and wind speed at 1 meter above ground (Wind). Although the meteorological dataset provides comprehensive coverage data for China from 1951 to 2021, it still contains gaps that cannot be inferred from remote sensing or other technical methods at either temporal or spatial scales. The data loss type, duration, and geographic distribution of these gaps are presented in Sect 2.1.2.

These essential records are extracted from the original meteorological dataset, they also are cleaned and unified to standard data format (Sect 2.3). All the original synoptic data have undergone strict quality controls, including assessing spatiotemporal consistency, identifying outliers, and correcting suspicious and erroneous data (Du et al., 2020). The spatial distribution in different climate regions of China for each site is detailed in Figure 1. The meteorological station is not evenly distributed in China. In the marginal tropical humid zone (MTH), north subtropical humid zone (NSH), and the warm temperate semi-humid zone (WTSH), the station density is relatively high and distributed evenly. In the other four climate regions, The meteorological stations in these areas are mainly distributed in areas with high human activity, such as the eastern part of the Plateau temperature semi-arid zone (PTSA), The western part of the Mid temperature Arid Zone (MTA), The southwestern part of the Mid temperature semi air zone (MTAS) and the areas other than the northern part of the Mid temperature semi air zone (MTSH).

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

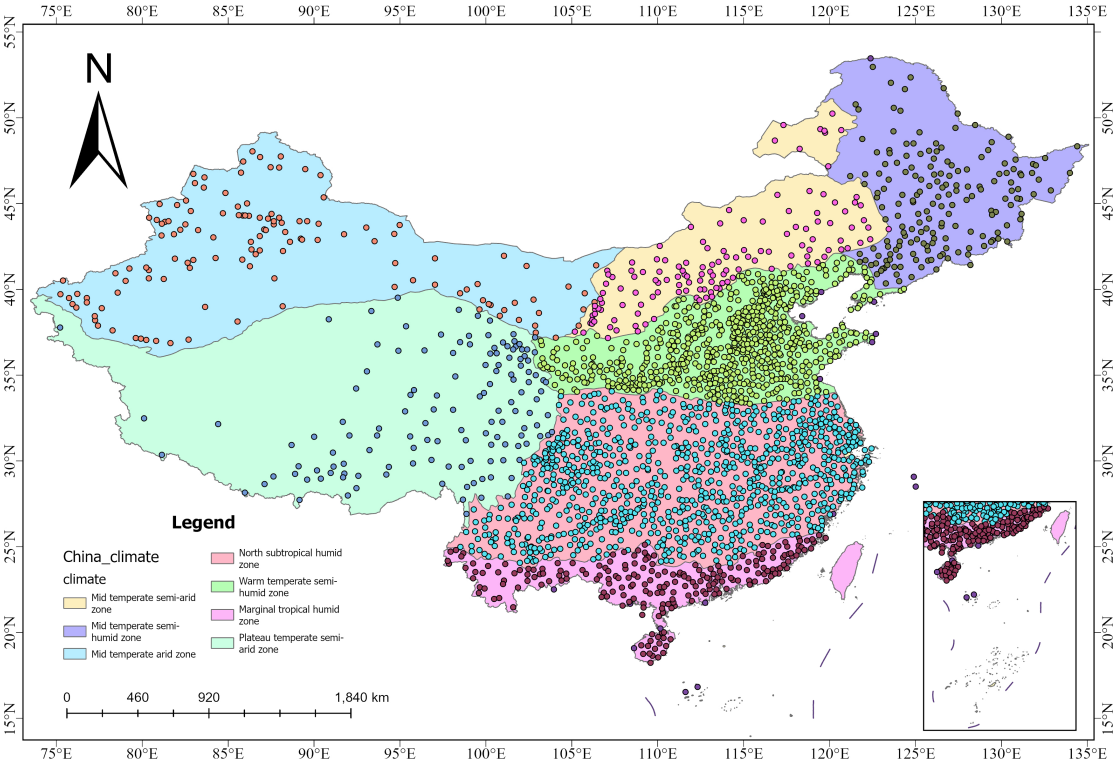


Figure 1: Spatial distribution of meteorological sites and climate zones aligns.

This approach allows us to derive the necessary solar radiation values indirectly, ensuring that the PM formula can still be applied effectively. By estimating Rs from SSD, we maintain the integrity and continuity of the ETo computations despite the absence of direct solar radiation measurements. This estimation incorporated the day of the year, calculated from the date in the records, and geographical coordinates, derived from longitude and latitude. Detailed methodologies and conversion formulas for this process are provided in Sect. 2.2.1. (Yang et al., 2006). While several datasets offered information on solar radiation for China (Tang et al., 2013), the derived solar radiation from the original dataset is more suitable for computation because it aligns well with the geographical location.

2.1.2 Data Gap Details

To quantify and classify these data losses, statics analysis was conducted for the meteorological dataset of the frequency, categories, and severity based on data indices' absence types. The spatial distribution of data loss for each site are presented in Fig. 2. The numerical analysis related to these observations are presented in Table 1. The analysis shows that 74 sites exhibit gaps ranging from 1,800 to 3,284 days (approximately 10 years), with 53 of these 74 sites having record gaps exceeding 10 years. Aside from these 74 outlier sites, data loss is evenly distributed across both temporal and geographic scales for all other sites. The average duration of the data loss period is 743 days, with a median of 465 days. These findings indicate that, while substantial data gaps exist across the dataset, the overall long-term reliability of the data remains robust. However, the problem

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

of data loss is also severe for a complex dataset because long-time data loss and specific climate variables loss exited and would hugely affect the ETo estimation quantity by machine learning models.

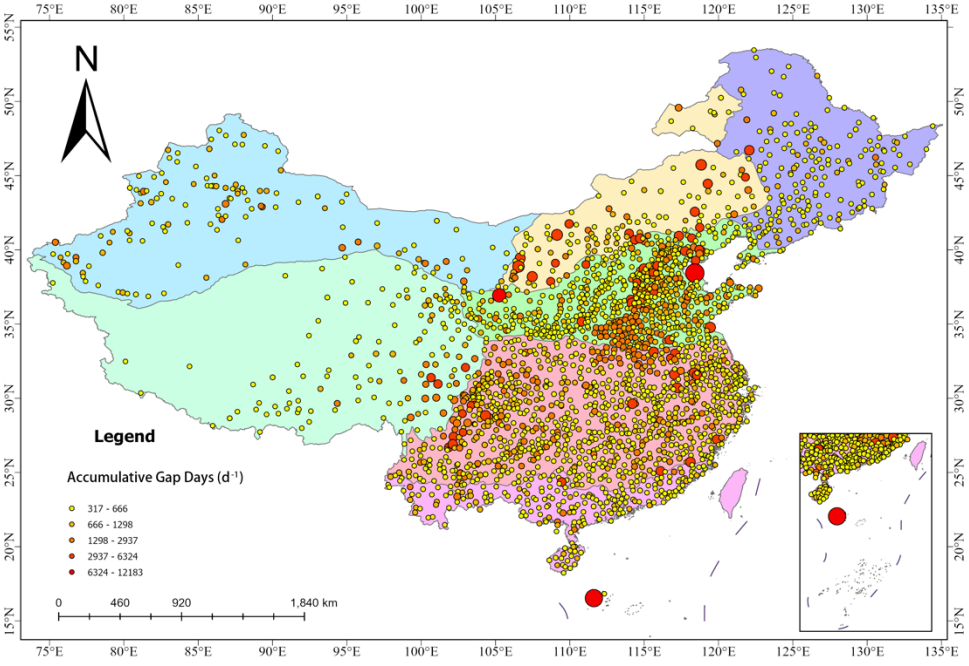


Figure 2: Spatial distribution of Gap days in each site.

Table 1: Record days, accumulative gap days, and gap percentage

	Record Days (d^{-1})	Gap Days (d^{-1})	Gap percentage（%）
Average	22626	743	3.43
Median	23128	465	2.17
Max	25933	12183	99.04
Std	3200	774	4.48
25%	22641	385	1.72

To facilitate machine learning models in learning features from incomplete data and filling the gaps in ETo, all instances of missing meteorological data were classified. The data gaps were categorized into 26 types (Table 2) based on the combinations of missing daily variables. The basic five data loss types (S1-S5) result from the absence of a single variable out of the total five, including maximum temperature loss (Tmax, S1), minimum temperature loss (Tmin, S2), sunshine duration loss (SSD, S3), wind speed at 1 meter above the ground loss (Wind, S4), and relative humidity loss (RH, S5). These five basic gap types account for more than 75% of the total data gaps in the entire dataset (Fig. 3). The detailed volume of each gap type is presented in Table 3. To simplify the expression, we will use the terms Tmax, Tmin, RH, SSD, Wind, SSD-Wind and Wind-RH to represent the data gap type S1-S7, where the referenced index indicates the missing variable. The other 21 gap types (S6-S27) are detailed in Table 2. According to Fig. 3, the data gap types S1-S7 cover more than 90% of the total data loss type among

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

all 2419 sites, and the remaining data of S6/S7 is also sufficient for machine learning models. Therefore, we extracted the SSD-Wind loss (S6) and RH-Wind loss (S7) for gap-filling and analysis using the same methods as for S1-S5. The quantity of the other 19 data gap types (S8-S26) is significantly less than that of data loss types S1-S7 (Table 3). Consequently, these data gaps are still filled using reliable LSTM models (Sec. 3.4).

Table 2: Gap type code

Gap Type of lost data	Code
Tmax loss	S1
Tmin loss	S2
SSD loss	S3
Wind loss	S4
RH loss	S5
SSD&Wind loss	S6
Wind&RH loss	S7
Tmax&Tmin loss	S8
Tmax&SSD loss	S9
Tmax&Wind loss	S10
Tmax&RH loss	S11
Tmax&Tmin&SSD loss	S12
Tmax&Tmin&Wind loss	S13
Tmax&Tmin&RH loss	S14
Tmax&SSD&RH loss	S15
Tmax&Tmin&SSD&RH loss	S16
Tmin&SSD loss	S17
Tmin&Wind loss	S18
Tmin&RH loss	S19
Tmin&SSD&Wind loss	S20
Tmin&SSD&RH loss	S21
Tmin&Wind&RH loss	S22
Tmin&SSD&Wind&RH loss	S23
SSD&RH loss	S24
SSD&Wind&RH loss	S25
Tmax&Tmin&SSD&Wind&RH loss	S26

The distribution of data loss quantity over the period from 1951 to 2021 is presented in Fig. 4(a). Data loss is primarily concentrated in three time periods: 1951-1975, 1985-2000, and 2020-2021, with the most severe data loss occurring in 2020-2021. Figure 4 (b) presents the total number of sites that contained a certain type of data gap in a year. As the total site quantity quickly increased from 1951 through 1960, this period included a large data gap of Wind loss and SSD loss. From 1960 to

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

1980, with the total loss quantity decreasing soon and reaching a low level in 1975, the SSD loss is more commonly seen in nearly half of the meteorological stations. Meanwhile, Wind loss began to increase in about half of the total meteorological sites and became the most widely distributed data loss type from 1975 to 2005. Between 2005 and 2019, both the quantity and the types of data loss decreased quickly. Still, the data loss became severe again in 2020 and 2021 because the SSD loss and wind loss appealed in all synoptic stations and the total data loss quantity increased quickly, refer to Fig.4 (b).

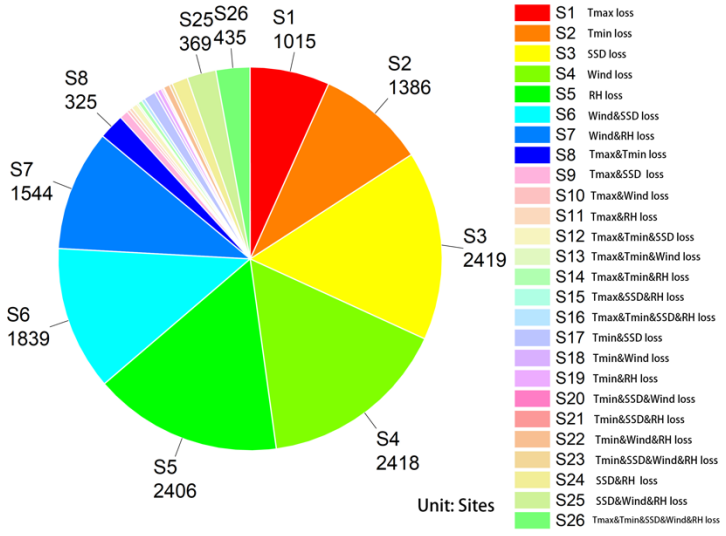


Figure 3: Meteorological station quantity for each data gap type.

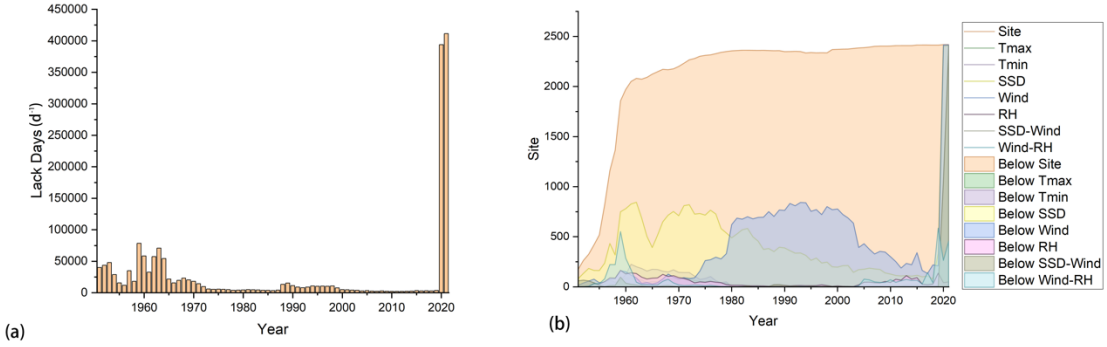


Figure 4: the gap volume of each year(a), the site quantity of each gap type in each year (b).

Figure 5 (a-g) displays the spatial distribution and histograms of the 7 major types (S1-S7) of data loss, the size of the dot represents the cumulative number of data loss quantity corresponding to each specific data loss type. It could be seen that the Wind Speed loss is the most severe of all types. The Wind Speed loss is evenly distributed in each meteorological station, this data loss type also produces long-time data loss across China. The numerical statistic results show that wind speed loss become severe since the 1980s and is the most severe problem for the synoptic data in 2020 and 2021. Other data loss types happened relatively less, and severe data loss only occurred in certain regions of China. Severe Tmax loss and SSD loss appealed mostly

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

in northern areas of China, while Tmin loss happened around the middle part of China. The parameters Tmax and SSD both offered important information for solar radiation, in arid regions, lacking one variable in these two parameters would not cause direct problems in ETo filling models. Severe RH loss occurred in the south-west region of China. SSD-Wind loss and Wind-RH loss distributed less than the original Wind loss, but still widespread across China.

Table 3 The gap quantity of each gap type

	Gap Days（d^{-1}）	Gap Percentage（%）	Site Mean（d^{-1}）	Site std（d^{-1}）	Site Percentage Mean（%）	Site Percentage Std (%)
Wind (S4)	896320	1.64	371	159.2	1.75	1.64
SSD (S3)	626709	1.15	264	663.6	1.21	3.75
Tmax (S1)	13988	0.03	13.8	50.7	0.02	0.14
Tmin (S2)	12689	0.02	9.16	37.73	0.02	0.12
RH (S5)	71822	0.13	46.5	322.6	0.13	1.15
SSD&Wind (S6)	85915	0.16	35.7	100.5	0.15	0.41
Wind&RH (S7)	14278	0.03	7.77	93.93	0.03	0.35
Others (S8-S27)	56345	0.1	0.02	0.01	0.08	0.12
Total	**1778066**	**3.25**	**747**	**178.6**	**3.39**	**1.1**

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

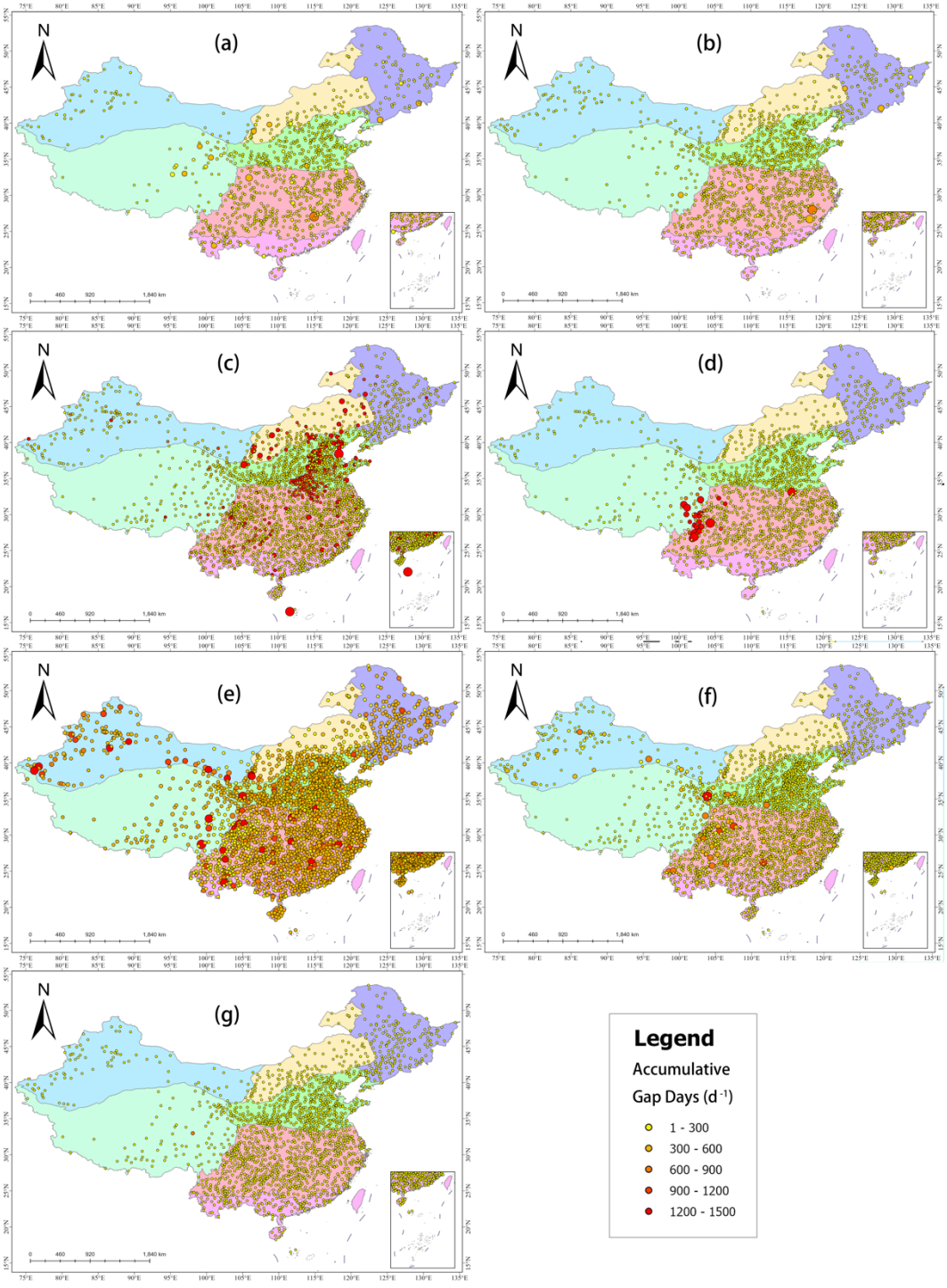


Figure 5: The quantity distribution of gap type Tmax (a), Tmin (b), SSD (c), RH (d), Wind (e), SSD-Wind (f), Wind-RH (g) for each site.

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

2.2 Methodology

2.2.1 ETo Calculation

The FAO-56 Penman-Monteith (PM) formula is used to calculate ETo. In consideration of both energy balance and aerodynamic terms, it offers a detailed estimation of the water demand of a well-watered grass field under prevailing environmental conditions. This formula is recommended by the Food and Agriculture Organization of the United Nations (FAO) within the FAO Irrigation and Drainage Paper No. 56. The FAO-56 PM formula is Eq. (1).

$$ET_0 = \frac{0.408\Delta(R_n - G) + r\frac{900}{T+273}u_2(e_s - e_a)}{\Delta + r(1 + 0.34u_2)} \quad (1)$$

Where Δ is the slope of the vapor pressure curve (kPa/℃), R_n is the net radiation at the crop surface (MJ m^{-2} d^{-1}), G is the soil heat flux density (MJ m^{-2} d^{-1}), r is the psychrometric constant (kPa/℃), T is the average daily air temperature at 2m height (℃), u_2 is the wind speed at 2 m height (m s^{-1}), e_s is the saturation vapor pressure (kPa), e_a is the actual vapor pressure (kPa), $(e_s - e_a)$ is the saturation vapor pressure deficit (kPa).

Another problem is that the PM formula required Rs, but the meteorological dataset only provided the sunshine duration (SSD). So, we use the Angström-Prescott equation in Eq. (2) to convert sunshine duration to solar radiation.

$$R_s = R_a\left(a + b\frac{S_0}{S}\right) \quad (2)$$

Where the R_s is the solar radiation (MJ m^{-2} d^{-1}), R_a is the standard solar radiation (MJ m^{-2} d^{-1}), S_0 is the monitored sunshine duration in a day (h d^{-1}), and S is the total daily sunshine hours (h d^{-1}), a and b are empirical coefficients that vary depending on the location and are determined from historical solar radiation and sunshine duration data.

2.2.2 ETo Calculation

The original meteorological dataset endured several standardization issues, including longitude and latitude changes due to site relocation or evacuation, the special code blended with the original data records, wrong measurement units, and other problems that occurred due to human errors. We deployed an automated program to recalibrate the standardization problems to minimize data fluctuations. We updated the geographical information to the newest records (2016) for each site, unified the units of data records, and aligned special meteorological information descriptions with the newest standards. This program is also used to detect and correct anomalies to change the different units that happened in data records. The data-cleaning process is essential for the calculation process.

Figure 6 is the flow chart of the whole ETo prediction and gap-filling process. After the data-cleaning process, the records without major index loss would straightly be used to calculate ETo with the FAO-56 PM formula. The data records with meteorological data gaps are extracted specifically and classified into the 26 data gap types in Sec. 2.1.2. The program would recognize the data loss type automatically and fill the ETo record according to the specific data loss type. To fill the gap data, the program firstly distinguishes whether the remaining climate data type is more than 3 or not; if there are still 3-4 environment parameters remaining, three gap-filling model based on the MARS/SVR/XGBoost algorithm would be trained and employed

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

to fill the ETo data records for the specific site and data loss type. If there are more than 2 parameters lost, the Long-short Term Memory algorithm is hired to create the gap-filling model. The gap data would be input to a specific trained LSTM model with the true ETo for the 7 previous days. Details of the pros and cons of the two methods are presented in Sect. 4.2 In short, the XGBoost model is more useful when there are only a few indices loss, and this approach could easily achieve high-quality prediction when continuous data are missing. The LSTM method could deal with the multiple parameters missing scenarios, especially when a whole day of environmental data are missing, and XGBoost could barely work. LSTM would give a relatively acceptable ETo prediction result.

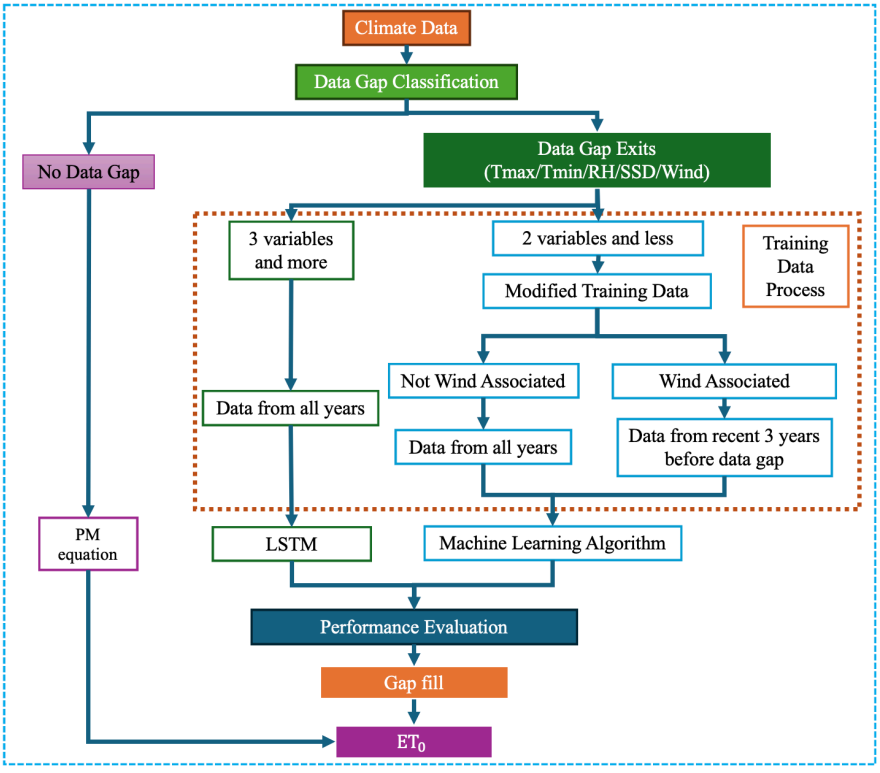


Figure 6: Flow chart of the gap-filling process.

Four models are introduced for the regression task for the 2 different data loss scenarios. The regression models could perform well in both continuity and discontinuity gaps in such scenarios. So, these models that ignore the relationship between data can be used for regression tasks in this scenario. In our research, the MARS (Multivariate Adaptive Regression Splines), SVR (Support Vector Regression), and XGBoost (Extreme Gradient Boosting) algorithms are used to simulate the single-day ETo from the remaining environmental indices.

MARS is a representative learning algorithm for non-parametric regression tasks because it is an ensemble of linear functions. It could be seen as an extension of linear models that automatically model nonlinearities and interactions between variables. This algorithm gradually classifies the data into more detailed classes by using a piecewise linear function composed of smaller functions like the “right function.” this basic function only determines whether the current value is greater than a

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

specific threshold. With the composition of all these basic functions, MARS could achieve good performance in high-deamination data classifying and regression tasks, and because the basic function part has a linear structure, the calculation of MARS is easier than other machine learning algorithms.

SVR is based on the concept of Support Vector Machines. It is a classical neural network algorithm. In SVR, the idea is to find a function that has at most a minus deviation from the obtained targets for all the training data and, at the same time, is as flat as possible. In regression tasks, this algorithm character could let it jump out of the smaller depression points, thereby enhancing the ability to resist outlier interference. This algorithm works well with both linear and non-linear data and is more robust against overfitting, especially in high-dimensional space. Compared to MARS, SVR has more advantages in dealing with dates that are not standardized enough.

XGBoost implements gradient-boosted decision trees (GBDT) designed for speed and performance. It is a scalable and accurate implementation of gradient boosting machines, one of the most powerful techniques for building predictive models. By learning from the gradient residual, XGBoost could handle more complicated data when there are large, larger internal differences. The metrological data are relatively stable in most regions. However, in regions where the weather changes fast and is easily influenced by big weather changes or human activities, the metrological data could be very challenging for learning algorithms because the outliner data might occur less regularly. The performance of these three 1-day data prediction methods is detailed in Sect. 3.2, which furtherly presents the comparison of the results of the three methods.

For the days that lack more than 2 environmental parameters, even those that lack records of this day, LSTM is a relatively reliable method to infer the ETo for the next day based on the previous 7 days' data records. LSTM, short of the Long Short-Term Memory algorithm, is a type of recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTMs have feedback connections that make them capable of processing not just single data points but entire sequences of data. The gates (input, output, and forget gates) allow LSTMs to selectively remember or forget patterns, which is crucial for understanding long-term dependencies in time series data and gives the LSTM model the ability to remember information for long periods.

LSTMs leverage the data from the previous 7 days to make accurate predictions for the next day, even without complete environmental parameters. By learning from the temporal patterns and relationships within the historical data, LSTMs can infer missing information and provide reliable ETo estimates. This capability is especially valuable when environmental data are incomplete or inconsistent.

2.2.3 Evaluation

To ensure the gap-filling result is reliable on both the statistical and absolute value levels, both the statistical and absolute value evaluation methods are introduced to estimate the prediction ETo. At the statistical level, R^2 (Coefficient of Determination), RMSE (Root Mean Squared Error), and nRMSE (Normalized Root Mean Squared Error) are the main indicators used to evaluate the reliability of the training data, validation data, and test data. More specific evaluation results are presented in Sect. 3.2 and 3.4. R^2 provides the measure of the strength and direction of the linear relationship between

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

observed and predicted values, indicating how well the independent variables explain the variance in the dependent variable. RMSE offers a measure of the differences between values predicted by a model and the values observed from the environment that is being modeled. It is used to quantify the model's accuracy in predicting the target variable on the same scale as the data. nRMSE is the RMSE normalized by the range or standard deviation of observed data, which allows for the comparison of models with different scales. The equation for R^2, RMSE and nRMSE are listed in Eq (3), Eq (4) and Eq (5).

$$R^2 = 1 - \left(\frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{c}\right) \quad (3)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (4)$$

$$nRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{\bar{y}} \quad (5)$$

In Eq (3), Eq (4) and Eq (5)., N represents the total observation sample size, and y_i *and* $\hat{y}_i$ is the i-th observation value and prediction value, $\bar{y}$ is the average of y.

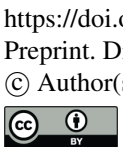
3 Result

3.1 ETo Dataset

A high-precision ETo evaporation dataset with the gap filled in the period of 195-2021 for mainland China is proposed. Figure 7 (a) and (b) shows the density distribution of mean daily evapotranspiration for the five statistic lines. The black and green lines almost coincided with the black lines, which are the density lines of daily mean ETo with/without the data gaps. The distribution is almost the same because the 3.6% data gap didn't heavily affect the overall trends, but there's still some difference between the two kinds of daily mean ETo (the purple line in Fig. 7 (b)). The average daily ETo with filled data are slightly lower than the daily ETo without filled data because the filled data tends to be lower than the original data (blue and red lines in Fig. 7 (b)). The main reason for the decline in average value is that the missing data didn't distribute evenly; the lower ETo range missed more data than the higher range and thus induced the lower density pick of the gap-filled results.

Figure 7 (c-i) displays the scatter plots of the predicted PM-56 ETo, and its true values with a dotted line represent the strictly equal scenario. When the parameters for the day only lost one of the indices in Tmax/Tmin/SSD/RH, it is clear to see that these 4 basic gap types were not influenced heavily by the data loss, and the R^2 for filled data are 0.98, the spatial distributions for each data gap type are detailed in Sect. 3.2. The data gaps associated with Wind, including Wind loss, SSD-wind loss, and RH-wind loss, showed worse simulation results compared to the 4 basic data gaps (Fig. 6 d, e, f). And there's a clear trend that the ETo tends to be overestimated when the parameter Wind is missing. This might indicate that the parameter Wind contributes significantly to daily ETo and has more impact in long periods and multiple areas. The same cases could be found in SSD-Wind missing, with the main range of estimation error being about 2 mm d^{-1}. When the parameter set Wind-RH is unavailable, the model tends to make errors in underestimating ETo. The bad results from the wind-associated computations

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024



might indicate that although a lot of research proves wind might not be a key parameter for ETo estimation, it still cannot be replaced by other parameters in a wide range.

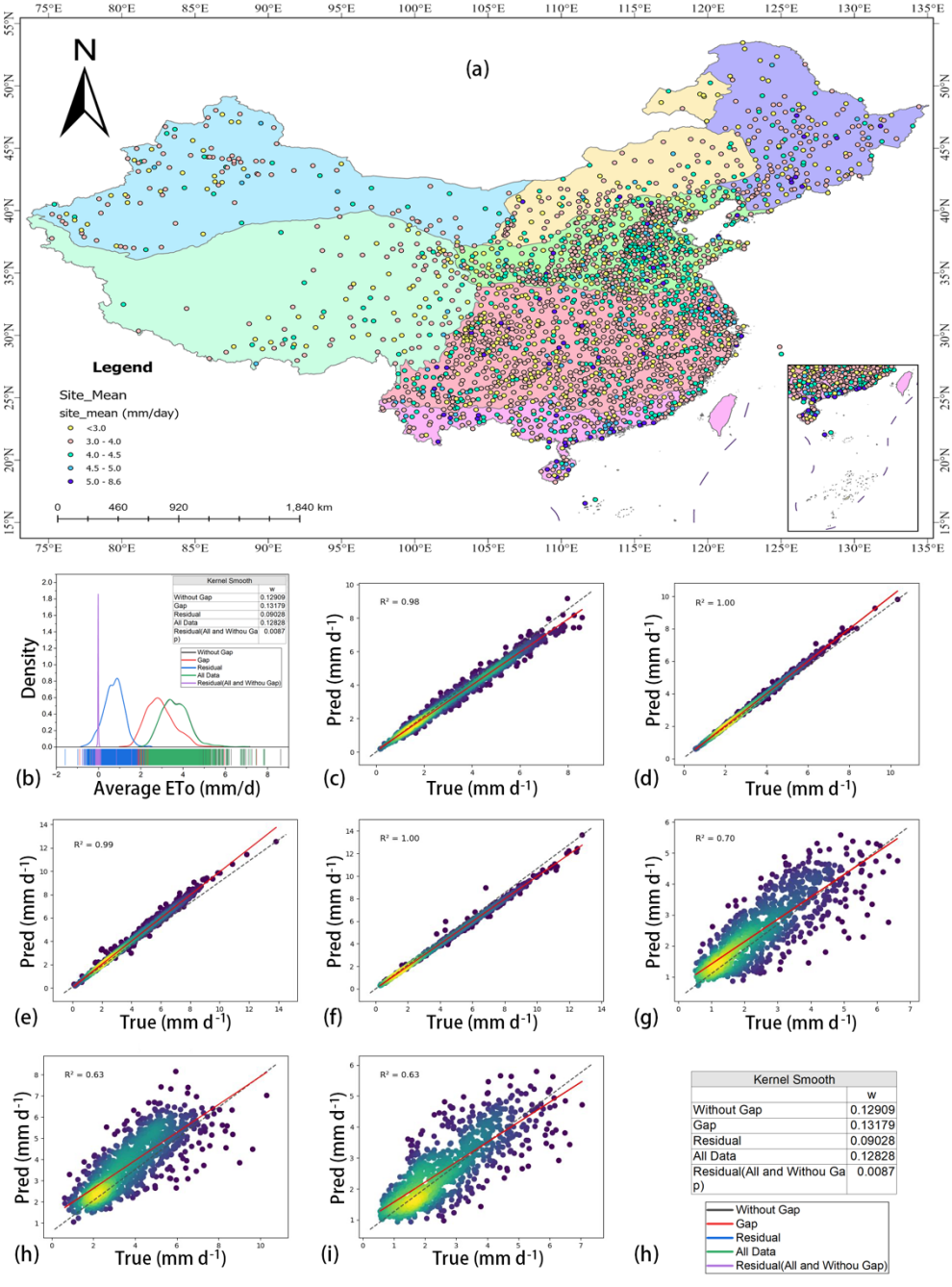
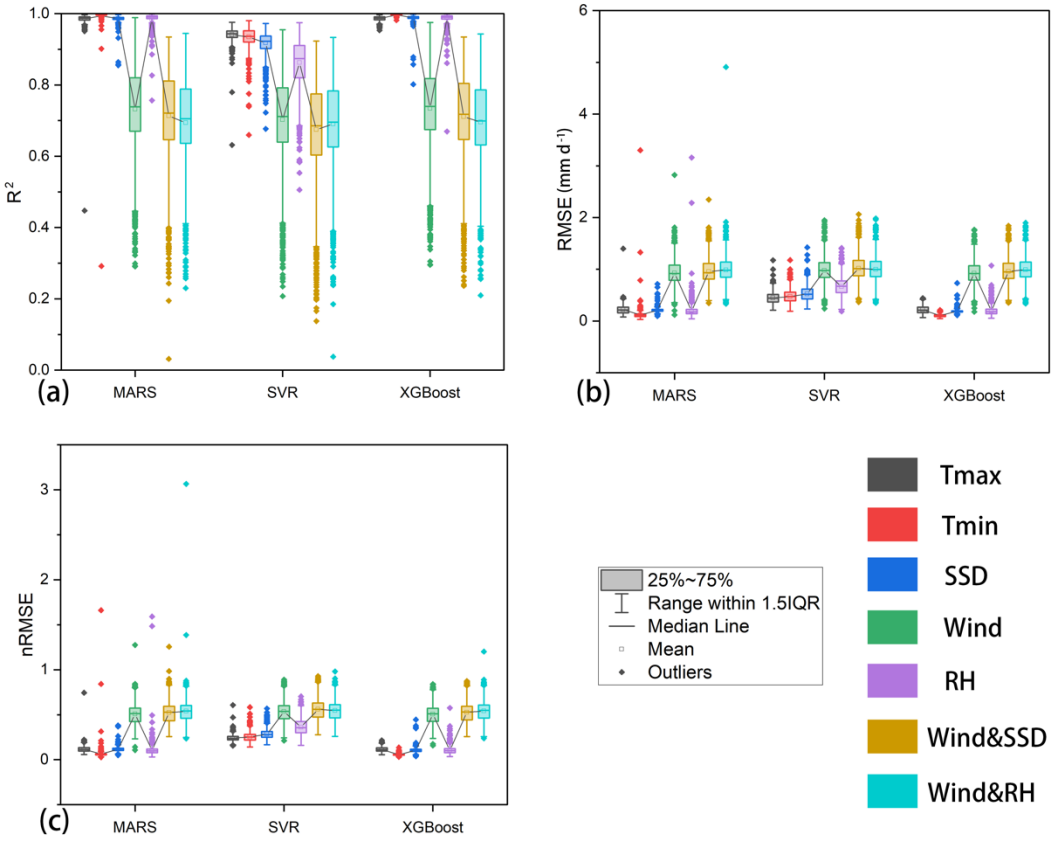


Figure 7: Daily ETo of 1951-2021 for each site (a), the daily ETo distribution of all sites (b), the scatter of true data and prediction ETo for each gap type including Tmax (c)/Tmin (d)/SSD (e)/RH (f)/Wind (g)/SSD-Wind (h)/Wind-RH (i).

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

3.2 Simulate Results



335 **Figure 8: Daily ETo of 1951-2021 for each site (a), the daily ETo distribution of all sites (b), the scatter of true data and prediction ETo for each gap type including Tmax (c)/Tmin (d)/SSD (e)/RH (f)/Wind (g)/SSD-Wind (h)/Wind-RH (i).**

Table 4: The statistic table of R^2, RMSE, and nRMSE for each gap type and machine learning algorithm

	MARS R2	XGBoost R2	SVR R2	MARS RMSE	XGBoost RMSE	SVR RMSE	MARS nRMSE	XGBoost nRMSE	SVR nRMSE
Tmax	0.99	0.99	0.9	0.22	0.22	0.45	0.12	0.11	0.24
Tmin	0.99	1	0.9	0.12	0.11	0.47	0.06	0.06	0.25
SSD	0.99	0.99	0.9	0.21	0.19	0.52	0.12	0.1	0.28
Wind	0.73	0.73	0.7	0.93	0.92	0.98	0.51	0.51	0.53
RH	0.99	0.99	0.9	0.19	0.19	0.66	0.1	0.1	0.36
Wind&SSD	0.71	0.71	0.7	0.95	0.96	1.02	0.52	0.53	0.56
Wind&RH	0.69	0.7	0.7	0.99	0.99	1	0.54	0.54	0.55

Four machine learning algorithms are applied to deal with different kinds of events in the meteorological dataset. We use three machine learning algorithms (MARS\SVR\XGBoost) to fill the gaps induced by single parameter loss, including 340 Tmax\Tmin\SSD\Wind\RH loss and the two major data loss types associated with wind (SSD-Wind loss, Rh-Wind loss).

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

Figure 8 presents the model capability of three machine learning methods for the seven types of environmental factors loss data. R^2, RMSE, and nRMSE are used to evaluate the regression fitting results for different models. Three machine learning methods performed well in the main 7 gap types. The R^2 for the four single parameter-missing types and three algorithms is very close. Table 4 shows the R^2, RMSE, and nRMSE for the MARS/XGBoost/SVR model for the major 7 gap types. Other than data gaps associated with the parameter Wind.

It could be seen that solitary parameter loss would not affect machine learning algorithm performance deeply except the Wind and the parameters missing associated with wind, including SSD-Wind and Wind-RH, the R^2, RMSE, and nRMSE are all decreased quickly compared to the solitary parameter missing. In most studies, the wind is not an important affective index in the input dataset for two main reasons. The first reason is that soil humidity is not enough for storage in arid and semi-arid areas, so wind speed is less affected than radiation heat. Another reason is that the relative humidity and solar radiation might not be enough to describe the true information for the land surface; the usage type of the land surface would also influence the evapotranspiration process by affecting wind speed.

It could be derived that different machine learning models would act differently in regression tasks. However, the evapotranspiration regression task might not be complex enough to induce significant differences for MARS, XGBoost, and SVR. The comparison result can be seen in Fig. 8. Though there are several differences in prediction results, no significant performance difference was displayed. A more detailed discussion of the difference between these three algorithms is presented in Sect. 4.2. Among these algorithms, XGBoost is the best performing algorithm, so the gap-filled results from XGBoost are chosen to be the result for 7 major gap types S1-S7.

Figure 9 (a-g) is the spatial distribution of the R^2/RMSE/nRMSE for all meteorological sites' 7 main gap types. The Marginal Tropical Humid Region showed a significant decrease compared to northern regions. It could be found that wind influences both simulation precision and the numerical value of the prediction result more than other regions. In the Marginal Tropical Humid Region of China, wind may be more highly related to both evaporation and perception processes than other regions. The commonality of the impact of wind speed on all stations in this climate zone may indicate that some simplified models are not entirely applicable in the Marginal Tropical Humid Region.

3.3 Large error

The precise prediction aims to estimate low numerical error volume results for parameter missing days, so the absolute error value must be within an acceptable range in the practical process. Thus, we estimated the amount of data with an absolute error greater than 1.5 mm d^{-1} for each data gap type in the warm temperature semi-humid region (WTSH), marginal tropical Humid region (MTH), north subtropical humid region (NSTH), plateau temperate semi-arid region (PTSA), mid temperate semi-humid region (MTSH), mid temperate semi-arid region (MTSA), mid temperate arid (MTA) as shown in Table 5. The quantity and type of data loss both affect the large error quantity for each climate zone.

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

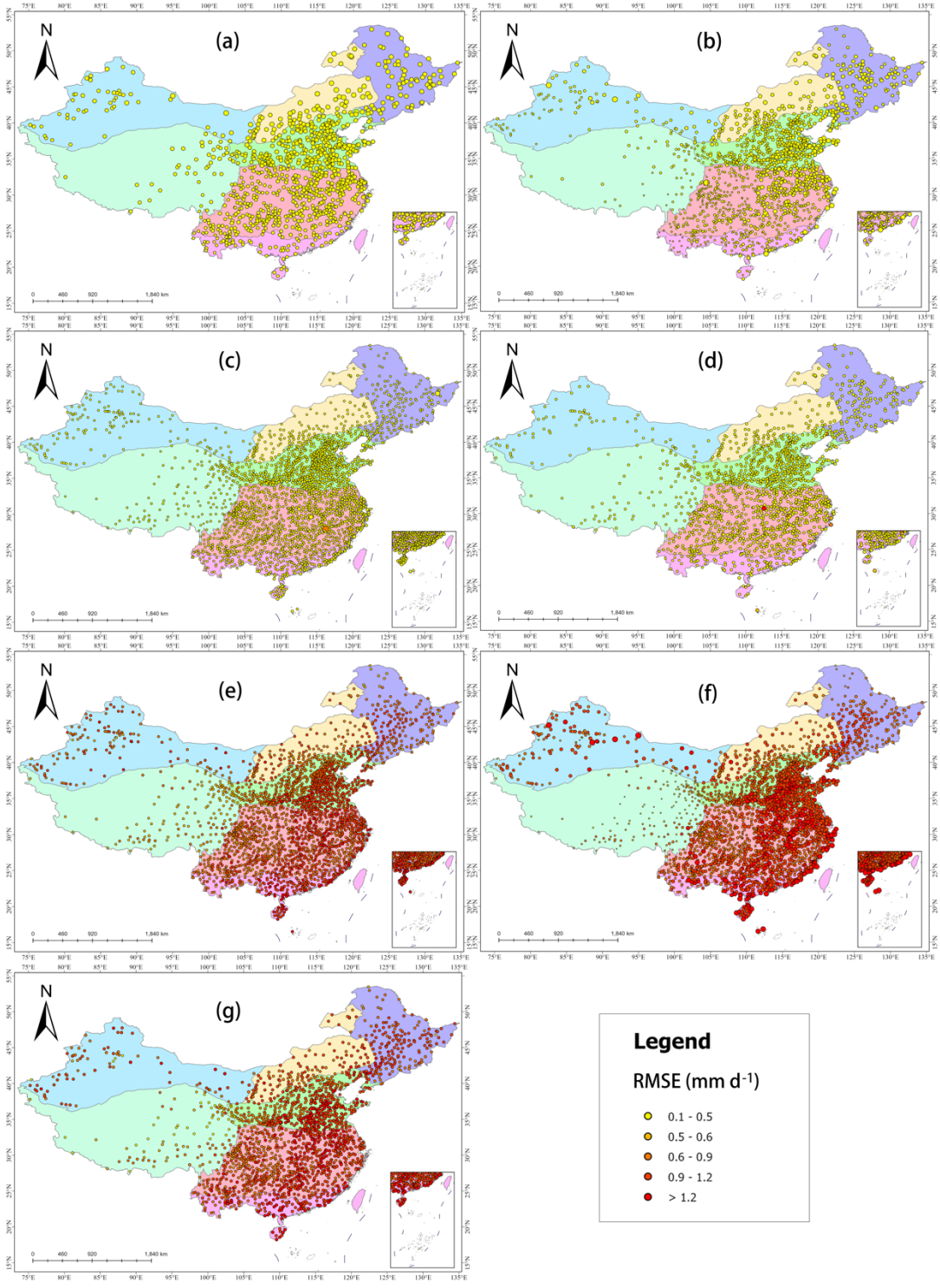


Figure 9: (a-g) The spatial distribution of RMSE for each gap type.

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

Except for the temperate and semi-arid areas on the plateau, the large error that the absolute error is greater than 1.5 mm d^{-1} is mainly caused by Wind-associated data gaps, and the quantity proportion of large error accounts for approximately 0.18% in the total data quantity. This means this dataset is reliable for the data quantity with severe error is few. From the distribution of large errors for different data gaps in the 7 climate regions of China, it could be derived that though there are not many large error results, there's still clearly a difference in large error quantity for different regions and gap types. Compared to absolute error quantity, in Semi-Arid Mid Temperate regions and Semi-Humidex Mid Temperate regions, there are more large errors caused by Tmax than in other regions. The large error caused by Tmin lost only is the least in all data gap types. The large errors caused by the parameter RH missing are less than SSD missing, but the contribution to the total number of large errors of these two data gaps is relatively small compared to the Wind-associated gaps. The data gaps associated with Wind contributed the most quantity of large errors due to two reasons; the first is the simulation model of these data gaps didn't return idealistic results, and the other reason is that the data gaps associated with Wind contributed a lot in the total data gap quantity (Sect. 2.1.2).

Table 5: Estimation quantity of the absolute error of ETo prediction > 1.5 mm d^{-1}

CLIMATE	TMAX (%)	TMIN (%)	SSD (%)	WIND (%)	RH (%)	SSD&WIND (%)	WIND&RH (%)
WTSH	0.00*	0.00*	0.00*	0.18	0.00*	0.02	0.00*
MTH	0.00*	0.00*	0.00*	0.21	0.00*	0.02	0.00*
NSTH	0.00*	0.00*	0.00*	0.18	0.00*	0.02	0.00*
PTSA	0.00*	0.00*	0.00*	0.03	0.00*	0.00*	0.00*
MTSH	0.00*	0.00*	0.00*	0.14	0.00*	0.01	0.00*
MTSA	0.00*	0.00*	0.00*	0.15	0.00*	0.01	0.00*
MTA	0.00*	0.00*	0.00*	0.19	0.00*	0.03	0.00*

* MEANS ACTUAL VALUE < 0.01

3.4 Multiple data-loss gaps

For data loss types S8-S26, we employed the Long Short-Term Memory (LSTM) network to fill gaps resulting from multiple data losses. These gaps constitute a minor portion of the total data volume (as detailed in Sect. 2.1.2) buts pose significant challenges for regression models due to insufficient remaining information for training. To estimate these data gaps, information from the preceding and succeeding ETo series should be utilized. The LSTM model uses inputs from the previous seven days to infer the ETo for the missing day. To aid the LSTM network in recognizing gaps in the data, we replaced the missing parameter with -1 as a placeholder in the input data and subsequently trained the entire neural network to predict the ETo for the next day. The data used for training, validation, and prediction at each site is the gap-filled data provided by the XGBoost model. The training data for each site consists of data excluding the validation and test datasets, where the validation data are from the years 2010-2019 and the test data are from the years 2020-2021. Based on the training the model for each site, the program fills the data gap by inputting the seven-day data preceding the gap, subsequently predicting the filling result for the data gap.

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

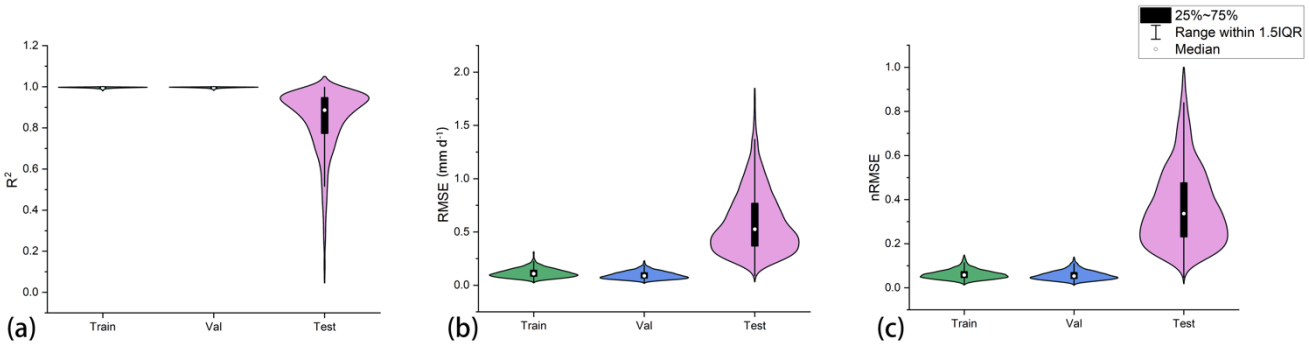


Figure 10: The LSTM algorithm performed well in most meteorological stations, and the gap-replacing methods could be used to fill the uncontained data gaps.

4 Discussion

4.1 Importance Distribution

The importance is derived from the model simulation process, and it could demonstrate the most important synoptic variables for ETo prediction under limited data (T. Hu & Song, 2019). The synoptic variables are highly entangled with each other according to the local environment condition, so when one or even more environmental variables are removed from input, the model could still predict ETo at high precision (proved in Sect. 3.2), and the importance for the synoptic variable under each data loss condition could straightly show the most entanglement with the missing data variable in different regions.

The geographical distribution of the mainly 7 types of data loss is presented in Fig. 11; the dot scale and colour represent the value and type for each site's max influence factor, all showing an importance value of more than 0.1 to exclude the random guess scenarios. There's a demarcation line between the north and south for Tmax loss. The main influence factor values are smaller in the South than in the North, which might indicate that the situation of synoptic variables coupling with Tmax is more complicated in the South than in the North. In the north of China, Tmax is highly entangled with Rs (Duo et al., 2016), and in the southwest region of China, Tmax might be entangled highly with the Wind Speed. The analogous phenomenon could be observed in the Tmin loss scenarios; in the south region of the North Subtropic Humid regions, Tmin has a strong relationship with Wind Speed. For SSD loss, Rs and Tmax take hold of most areas. The SSD is the source of Rs and is thus highly entangled with Tmax. For RH loss, the importance distribution is closer to Tmin loss; this might be because both the Tmin and RH data are highly influenced by ground information, so there's a stronger relation between these two synoptic variables. The last 3 data loss types associated with Wind, including Wind loss, SSD-Wind loss, and RH-Wind loss, have similar importance distribution of highly rely on the Rs and Tmax, which might indicate that the Rs and the Tmax change induced by Rs change is the most important driven force for evapotranspiration. This phenomenon according to the research Rs is the most driving variable for the evaporation process.

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

Nevertheless, at the same time, the huge variance of R^2 for data lacking Wind also displays the impact of wind speed and ground information in most areas of China. It is worth noticing that the influence of Wind is more important and significant in our paper than the conclusions derived from the other recent studies (Wu et al., 2019; Yin et al., 2020). These experiments proved that the Wind Speed is removable from the input indices set when using machine learning algorithms to predict ETo. To further address this conflict, we conducted a study to test how the different data years affect the simulation result and found that the newer the data year, the closer the fitting results are to the results stated in the current research. It could be found in the research that wind trends have continued to decrease across China since 1985, which might be the reason wind speed has taken a less important part in evaporation in simulation in recent years. This result showed a significant character change in the Wind Speed in ETo computation during the past few years. The research from (Liu & Zhang, 2013) showed that the wind speed became less important in ETo computation. However, the same phenomenon could not be found in other indices, which proved to be more accurate when the training data increased.

4.2 Comparing and uncertainties

Using the PM-56 formula to estimate ETo has largely proved to be a reliable approach, and machine learning algorithms have also proved to be a reliable way to predict ETo under limited data. In the past few years, lots of research has compared the different models and different combinations of variant input environment parameters as model input and has proved that machine learning algorithms could derive more precise information from huge data records than empirical equations. Our research got a homogeneous conclusion but also found out that the data record length and the outliner points might largely affect the model simulation results under limited data.

In training, the dataset covered all the scenarios that happened in the test or product area and was relatively flattened; the model would perform well, and both Mars\SVR\XGBoost could achieve high performance in predicting ETo. However, when dealing with high variance data, MARS tends to classify the data range with the segmentation strategy, resulting in the high sensitivity to outrange data, and the outrange data would take a high and useless decision tree to mislead the prediction. The SVR algorithm performed better in some situations than MARS but also endured the deficits that come with the classifying idea of the original algorithm. As an improvement to the classical algorithm, XGBoost does not use the original data to do the segmentation but to learn the residual to describe the data change, which directly allows XGBoost to perform well with more unreasonable values appealed in the training dataset. Research from multiple regions (Hossein Kazemi et al., 2020) and ours all show the XGBoost could deal with the outliner better than other algorithms benefiting from residual learning.

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

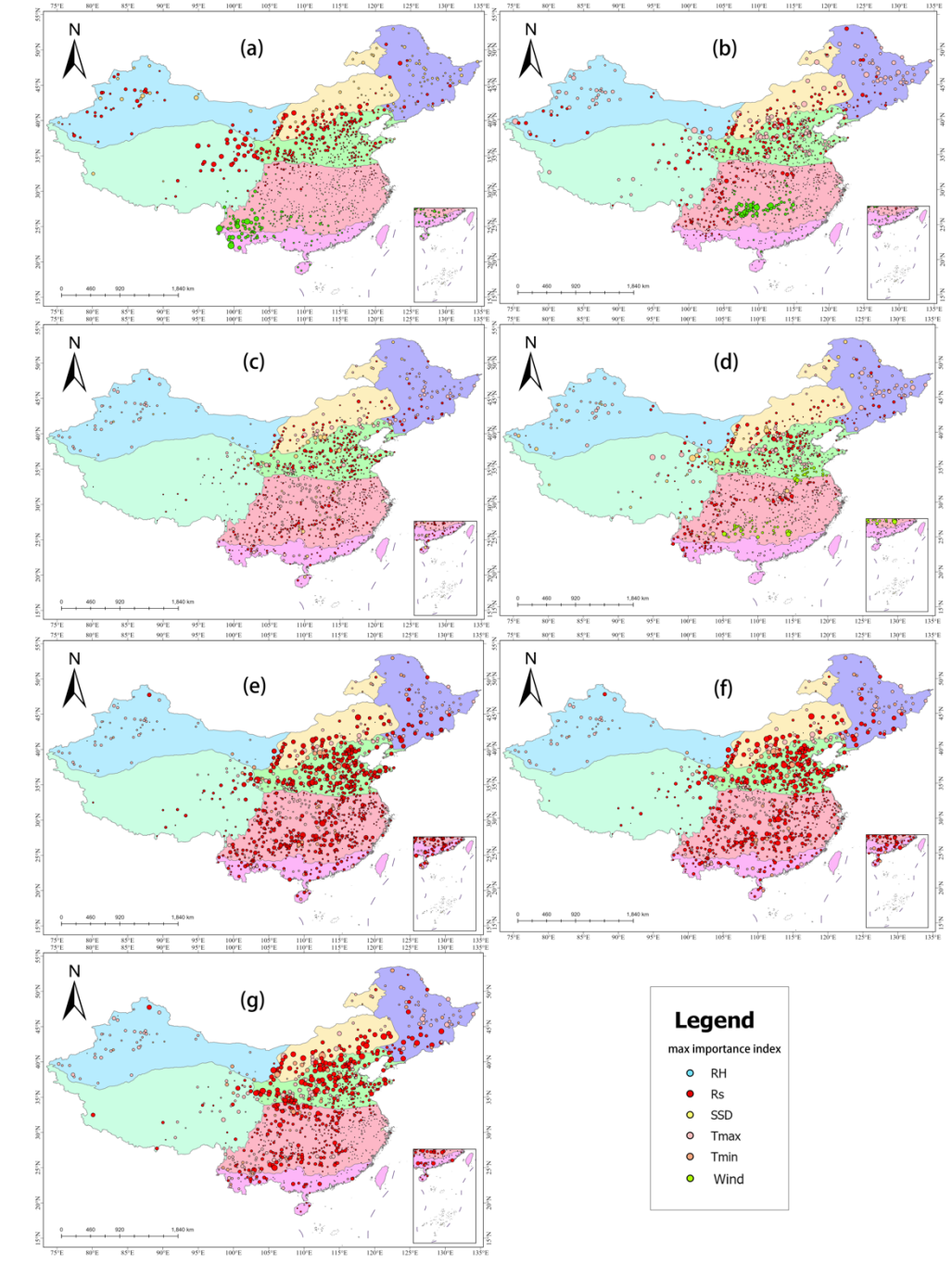


Figure 11: The spatial distribution of importance for (a) Tmax, (b) Tmin, (c) SSD, (d) RH, (e) Wind, (f) SSD-Wind, (g) Wind-RH loss type.

455

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

The time length for the input data would also affect the training result. In general, about 3 years would be sufficient for ETo prediction of a precise day or even 1-7 days (Yin et al., 2020). However, the discussion for time length about how long and how much environmental data should be covered for training is still few. Usually, the training data should contain all the synoptic scenarios so it could classify the test data, but the weather not only changes on a small scale but also varies a lot on a long-time scale. Our research found that in different indices loss scenarios, the longer temporal length led to better RMSE performance in the Tmax/Tmin/SSD/RH loss scene, but in parameters associated with Wind, the model quickly deteriorated when the year was prolonged. The reason for this phenomenon tends to be the wind speed at 2m could quickly affect the evaporation speed, but this variable could also be quickly controlled by the ground usage type or other human activity compared to other large-scale environmental indices. The ground usage type is easily changed and largely scale ground usage changing would directly lead to the wind speed change at 2m. Wind speed has been coupled strongly with air temperature and humidity; the results show that the wind variable has become less important in ETo prediction in recent years, which might indicate that the relationship between temperature, humidity, and wind speed has become closer than in years before. This might be another result of global warming for the ability to regulate climate is weakened, and the wind speed, which originally affects temperature and humidity, is instead affected by temperature and humidity.

In research for the past 10 years, LSTM algorithms have been proven to be a more reliable way to predict ETo under limited conditions (Rajput et al., 2023). Compared to the machine learning algorithms, LSTM algorithms performed better in deriving information from the continued data; thus, with actual ETo data from the previous day and the environmental parameters, the algorithms could output precise ETo estimation for the gap day. However, due to this algorithm's character, the LSTM algorithm is limited by the continuance of data. Although the gaps could be ignored during the training process, the LSTM can't output long prediction series without large errors. The longer the series is, the huger the errors might be. This problem couldn't be overcome by longer input, and the major 7 data gap types (S1-S7) contained long period data gaps, so we didn't use LSTM to fill the gap types S1-S7.

4.3 Applications of ETo dataset in the future

By filling the data gaps, we now present a complex ETo dataset for the 2419 synoptic stations of China across 1951-2021 at daily scale. Differing from the satellite products, this dataset contained data from 1951-1982 and presented the ETo from the ground synoptic stations with much more accuracy.

This ETo dataset has significant potential for various agricultural applications. The FAO-56 PM equation is calculated from a hypothetical field with standard grass cover and an ideal soil condition (L. S. Pereira et al., 2015). Still, the concept of reference evapotranspiration (ETo) is pivotal in managing and planning for the efficient use of water resources. For example, the true evapotranspiration (ET) could be estimated through ETo and easily facilitate the supervising and calibration programs. In agriculture, ETo is integral for achieving precision farming, identifying water stress areas, optimizing irrigation schedules, and adjusting irrigation volumes to the specific needs of crops in different regions around China. Combined with the eddy flux

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

data from flux tower sites, this ETo dataset could also be used to appropriately estimate the water cycle around China and the large trends of ET change during the past 70 years.

Although the programs and models are designed to fill the data gaps around China, they could also be widely used in similar scenarios. It is worth noticing that although the volume of monitoring data are large, it might still be too sparse for the country at both spatial and time scales. In recent years, the 3D neural network, which combined LSTM and the graph neural network, holds the most frontier position in the research of synoptic data. The Pangu model developed by Huawei Inc. (Bi et al., 2023) and the model from Deep Mind Lab (Lam et al., 2023) that use 3D sphere model to enhance the model performance using both spatial and temporal information. A complex dataset could facilitate the training process for these big climate models.

Overall, our dataset filled the data gaps by classifying and interpolating the data using the proper site-suited machine learning model. Machine learning and LSTM methods were reasonably used to process the missing data, resulting in a reliable ETo database based on the length of time and characteristics of the data gaps. Other studies can use more hydrological datasets with our ETo dataset to conduct further research on the long-term and large-scale ETo change, the actual evapotranspiration, the land-atmosphere water cycle, the big model for climate and other research directions for China with this dataset.

5 Data availability

The high precision gap-filled daily ETo data for China is archived and available at https://doi.org/10.5281/zenodo.11496932 (Zhou et al., 2024).

6 Conclusion

Based on the meteorological data provided by the National Climatic Centre of the China Meteorological Administration (NCC-CMA), we derived a high-precision gap-filled ETo dataset for mainland China using the FAO-56 PM formula. This dataset contained the meteorological data and the reference evapotranspiration data (ETo) for agriculture of 2419 sites across the period of 1951 to 2021, filling the gap in China's historical ETo data and exploring the machine learning algorithm performance in ETo prediction under limited data scenarios. To fill the gaps caused by long-period record changes and inevitable equipment destruction, we distinguished and classified the gap type by different data missing types, created the program for automatically extract suitable training data from the original dataset, and then used 3 different machine learning model to generate the suitable model for each data-missing less than 2 types in each site of the total 2419. As for the serious climate variable loss of more than 3, we employed the LSTM network to fill these types of gaps. The gap-filling methods achieved good gap-filling performance. that the XGBoost model achieved the best accuracy in all 3 machine learning models with high statistic levels. For the other 19 types of data gaps, the LSTM models were trained separately for each site and achieved average R^2, RMSE, and nRMSE at 0.9, 0.5 mm d^{-1}, and 38% for the total 2419 stations, and the absolute error of more than 1.5 mm d^{-1} are proved to be under 1% in the whole dataset.

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

Also, our result indicates that climate variable gaps existing in climate datasets would not directly lead to serious ETo data gaps because now the machine learning models could preserve the most data features for ETo change and come out with reasonable ETo predictions. This suggests that there's underlying patterns for ETo in most regions of China. Furtherly, the analysis of importance for different data gap models also reveals that the entanglement phenomenon between climate factors in different regions might could be explored through removing the synoptic variables from the input data series. The importance result shows although the Rs and Tmax induced by Rs is the most driving force for ETo, the Wind Speed is also the most important ground variables in evapotranspiration process. But combining results from other ETo research, the impact of Wind Speed might be continuously diminishing in recent years.

Author contributions

ZNS and WLF designed the research. ZNS, WLF, YQL, DJH developed the approaches and datasets. ZNS, WLF, YQL, Yang L, DJH, Yue L contributed to the analysis of the results and the writing of the paper.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgements

This work was financially supported by the Key Projects of Yunnan Provincial Department of Science and Technology (N0.202201AS070034), Key Laboratories of Yunnan Provincial Universities (KKPS201923009), and Key Projects of Yunnan Provincial Department of Science and Technology (No.202305AM070006). We especially thank all research subjects for their assistance participation in this study.

Financial support

This work was financially supported by the Key Projects of Yunnan Provincial Department of Science and Technology (N0.202201AS070034), Key Laboratories of Yunnan Provincial Universities (KKPS201923009), and Key Projects of Yunnan Provincial Department of Science and Technology (No.202305AM070006).

References

Allen, R. G., and Pruitt, W. O.: Rational Use of The FAO Blaney-Criddle Formula. Journal of Irrigation and Drainage Engineering, 112, 139–155, https://doi.org/10.1061/(ASCE)0733-9437(1986)112:2(139), 1986.

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

Bai, P., and Liu, X.: Intercomparison and evaluation of three global high-resolution evapotranspiration products across China. Journal of Hydrology, 566, 743–755, https://doi.org/10.1016/j.jhydrol.2018.09.065, 2018.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks. Nature, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Bormann, H., Diekkrüger, B., & Richter, O.: Effects of data availability on estimation of evapotranspiration. Physics and Chemistry of the Earth, 21, 171–175, https://doi.org/10.1016/S0079-1946(97)85580-2, 1996.

CHANG, J.-H.: AN EVALUATION OF THE 1948 THORNTHWAITE CLASSIFICATION. Annals of the Association of American Geographers, https://www.tandfonline.com/doi/abs/10.1111/j.1467-8306.1959.tb01594.x, 1959.

Chen, J. M., & Liu, J.: Evolution of evapotranspiration models using thermal and shortwave remote sensing data. Remote sSensing of Environment, 237, 111594, https://doi.org/10.1016/j.rse.2019.111594, 2020.

Chia, M. Y., Huang, Y. F., & Koo, C. H.: Support vector machine enhanced empirical reference evapotranspiration estimation with limited meteorological parameters. Computers and Electronics in Agriculture, 175, 105577, https://doi.org/10.1016/j.compag.2020.105577, 2020.

Du, J., Wang, K., Cui, B., & Jiang, S.: Correction of Inhomogeneities in Observed Land Surface Temperatures over China. Journal of Climate, 33, 8885–8902, https://doi.org/10.1175/JCLI-D-19-0521.1, 2020.

Duo, A., Zhao, W. J., Qu, X. Y., Ran, J., & Xiong, K.: Spatio-temporal variation of vegetation coverage and its response to climate change in North China plain in the last 33 years. International Journal of Applied Earth Observation and Geoinformation, 53, 103–117, https://doi.org/10.1016/j.jag.2016.08.008, 2016.

Elhaddad, A., & Garcia, L. A.: Surface Energy Balance-Based Model for Estimating Evapotranspiration Taking into Account Spatial Variability in Weather. Journal of Irrigation and Drainage Engineering, 134, 681–689, https://doi.org/10.1061/(ASCE)0733-9437(2008)134:6(681), 2008.

Fan, Z.-X., & Thomas, A.: Spatiotemporal variability of reference evapotranspiration and its contributing climatic factors in Yunnan Province, SW China, 1961–2004. Climatic Change, 116, 309–325, https://doi.org/10.1007/s10584-012-0479-4, 2013.

Gavilán, P., & Castillo-Llanque, F.: Estimating reference evapotranspiration with atmometers in a semiarid environment. Agricultural Water Management, 96, 465–472, https://doi.org/10.1016/j.agwat.2008.09.011, 2009.

Ghilain, N., Arboleda, A., & Gellens-Meulenberghs, F.: Evapotranspiration modelling at large scale using near-real time MSG SEVIRI derived data. Hydrology and Earth System Sciences, 15, 771–786, https://doi.org/10.5194/hess-15-771-2011, 2011.

Gocic, M., Petković, D., Shamshirband, S., & Kamsin, A.: Comparative analysis of reference evapotranspiration equations modelling by extreme learning machine. Computers and Electronics in Agriculture, 127, 56–63, https://doi.org/10.1016/j.compag.2016.05.017, 2016.

Gowda, P. H., Chavez, J. L., Colaizzi, P. D., Evett, S. R., Howell, T. A., & Tolk, J. A.: ET mapping for agricultural water management: Present status and challenges. Irrigation Science, 26, 223–237, https://doi.org/10.1007/s00271-007-0088-6, 2008.

Granata, F.: Evapotranspiration evaluation models based on machine learning algorithms—A comparative study. Agricultural Water Management, 217, 303–315, https://doi.org/10.1016/j.agwat.2019.03.015, 2019.

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

Guven, A., Aytek, A., Yuce, M. I., & Aksoy, H.: Genetic Programming-Based Empirical Model for Daily Reference Evapotranspiration Estimation. CLEAN – Soil, Air, Water, 36s, 905–912, https://doi.org/10.1002/clen.200800009, 2008.

Hargreaves, G. H., & Allen, R. G.: History and Evaluation of Hargreaves Evapotranspiration Equation. Journal of Irrigation and Drainage Engineering, 129, 53–63, https://doi.org/10.1061/(ASCE)0733-9437(2003)129:1(53), 2003.

Hossein Kazemi, M., Shiri, J., Marti, P., & Majnooni-Heris, A.: Assessing temporal data partitioning scenarios for estimating reference evapotranspiration with machine learning techniques in arid regions. Journal of Hydrology, 590, 125252, https://doi.org/10.1016/j.jhydrol.2020.125252, 2020.

Hu, T., & Song, T.: Research on XGboost academic forecasting and analysis modelling. Journal of Physics: Conference Series, 1324, 012091, https://doi.org/10.1088/1742-6596/1324/1/012091, 2019.

Hu, Z., Bashir, R. N., Rehman, A. U., Iqbal, S. I., Shahid, M. M. A., & Xu, T.: Machine Learning Based Prediction of Reference Evapotranspiration (ETo) Using IoT. IEEE Access, 10, 70526–70540, https://doi.org/10.1109/ACCESS.2022.3187528, 2022.

Huang, J.-C., Ko, K.-M., Shu, M.-H., & Hsu, B.-M.: Application and comparison of several machine learning algorithms and their integration models in regression problems. Neural Computing and Applications, 32, 5461–5469, https://doi.org/10.1007/s00521-019-04644-5, 2020.

Kim, S.-J., Bae, S.-J., & Jang, M.-W.: Linear Regression Machine Learning Algorithms for Estimating Reference Evapotranspiration Using Limited Climate Data. Sustainability, 14, 11674, https://doi.org/10.3390/su141811674, 2022.

Kinaneva, D., Hristov, G., Kyuchukov, P., Georgiev, G., Zahariev, P., & Daskalov, R.: Machine Learning Algorithms for Regression Analysis and Predictions of Numerical Data. 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 1–6, https://doi.org/10.1109/HORA52670.2021.9461298, 2021.

Kisi, O. : The potential of different ANN techniques in evapotranspiration modelling. Hydrological Processes, 22, 2449–2460, https://doi.org/10.1002/hyp.6837, 2021.

Kisi, O.: Modelling reference evapotranspiration using three different heuristic regression approaches. Agricultural Water Management, 169, 162–172, https://doi.org/10.1016/j.agwat.2016.02.026, 2016.

Kra, E. Y.: An Empirical Simplification of the Temperature Penman-Monteith Model for the Tropics. Journal of Agricultural Science, 2(1), p162, https://doi.org/10.5539/jas.v2n1p162, 2010.

Ladlani, I., Houichi, L., Djemili, L., Heddam, S., & Belouz, K.: Estimation of Daily Reference Evapotranspiration (ET0) in the North of Algeria Using Adaptive Neuro-Fuzzy Inference System (ANFIS) and Multiple Linear Regression (MLR) Models: A Comparative Study. Arabian Journal for Science and Engineering, 39, 5959–5969, https://doi.org/10.1007/s13369-014-1151-2, 2014.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., & Battaglia, P.: Learning skillful medium-range global weather forecasting. Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

LHOMME, J.-P.: A THEORETICAL BASIS FOR THE PRIESTLEY-TAYLOR COEFFICIENT. Boundary-Layer Meteorology, 82, 179–191, https://doi.org/10.1023/A:1000281114105, 1997.

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

Althoff, D., Dias, S. H. B., Filgueiras, R., & Rodrigues, L. N.: ETo-Brazil: a daily gridded reference evapotranspiration data set for Brazil (2000–2018), *Water Resources Research*, *56*, e2020WR027562, https://doi.org/10.1029/2020WR027562, 2020.

Li, Z., Zheng, F.-L., & Liu, W.-Z. Spatiotemporal characteristics of reference evapotranspiration during 1961–2009 and its projected changes during 2011–2099 on the Loess Plateau of China. Agricultural and Forest Meteorology, 154, 147–155, https://doi.org/10.1016/j.agrformet.2011.10.019, 2012.

Liu, X., & Zhang, D. Trend analysis of reference evapotranspiration in Northwest China: The roles of changing wind speed and surface air temperature. Hydrological Processes, 27, 3941–3948, https://doi.org/10.1002/hyp.9527, 2013.

Malik, A., Jamei, M., Ali, M., Prasad, R., Karbasi, M., & Yaseen, Z. M.: Multi-step daily forecasting of reference evapotranspiration for different climates of India: A modern multivariate complementary technique reinforced with ridge regression feature selection. Agricultural Water Management, 272, 107812, https://doi.org/10.1016/j.agwat.2022.107812j, 2013.

Mallikarjuna, P., Jyothy, S. A., Murthy, D. S., & Reddy, K. C.: Performance of Recalibrated Equations for the Estimation of Daily Reference Evapotranspiration. Water Resources Management, 28, 4513–4535, https://doi.org/10.1007/s11269-014-0733-9, 2014.

Mostafa, R. R., Kisi, O., Adnan, R. M., Sadeghifar, T., & Kuriqi, A.: Modeling Potential Evapotranspiration by Improved Machine Learning Methods Using Limited Climatic Data. Water, 15, Article 3, https://doi.org/10.3390/w15030486, 2023, 2023.

Paredes, P., & Pereira, L. S.: Computing FAO56 reference grass evapotranspiration PM-ETo from temperature with focus on solar radiation. Agricultural Water Management, 215, 86–102, https://doi.org/10.1016/j.agwat.2018.12.014, 2018.

Pereira, A. R., & Pruitt, W. O.: Adaptation of the Thornthwaite scheme for estimating daily reference evapotranspiration. Agricultural Water Management, 66, 251–257, https://doi.org/10.1016/j.agwat.2003.11.003, 2004.

Pereira, L. S., Allen, R. G., Smith, M., & Raes, D.: Crop evapotranspiration estimation with FAO56: Past and future. Agricultural Water Management, 147, 4–20, https://doi.org/10.1016/j.agwat.2014.07.031, 2015.

Rajput, J., Singh, M., Lal, K., Khanna, M., Sarangi, A., Mukherjee, J., & Singh, S.: Data-driven reference evapotranspiration (ET0) estimation: A comparative study of regression and machine learning techniques. Environment, Development and Sustainability, https://doi.org/10.1007/s10668-023-03978-4, 2023.

Roy, D. K.: Long Short-Term Memory Networks to Predict One-Step Ahead Reference Evapotranspiration in a Subtropical Climatic Zone. Environmental Processes, 8, 911–941, https://doi.org/10.1007/s40710-021-00512-4. 2021.

Salahudin, H., Shoaib, M., Albano, R., Inam Baig, M. A., Hammad, M., Raza, A., Akhtar, A., & Ali, M. U.: Using Ensembles of Machine Learning Techniques to Predict Reference Evapotranspiration (ET0) Using Limited Meteorological Data. Hydrology, 10, Article 8, https://doi.org/10.3390/hydrology10080169, 2023.

Santos, P. A. B. D., Schwerz, F., Carvalho, L. G. D., Baptista, V. B. D. S., Marin, D. B., Ferraz, G. A. E. S., Rossi, G., Conti, L., & Bambi, G.: Machine Learning and Conventional Methods for Reference Evapotranspiration Estimation Using Limited-Climatic-Data Scenarios. Agronomy, 13, 2366, https://doi.org/10.3390/agronomy13092366, 2023.

https://doi.org/10.5194/essd-2024-229
Preprint. Discussion started: 22 July 2024

Shamshirband, S., Amirmojahedi, M., Gocić, M., Akib, S., Petković, D., Piri, J., & Trajkovic, S.: Estimation of Reference Evapotranspiration Using Neural Networks and Cuckoo Search Algorithm. Journal of Irrigation and Drainage Engineering, 142, 04015044, https://doi.org/10.1061/(ASCE)IR.1943-4774.0000949, 2016.

Shiri, J.: Evaluation of FAO56-PM, empirical, semi-empirical and gene expression programming approaches for estimating daily reference evapotranspiration in hyper-arid regions of Iran. Agricultural Water Management, 188, 101–114, https://doi.org/10.1016/j.agwat.2017.04.009, 2017.

Tang, W., Yang, K., Qin, J., & Min, M.: Development of a 50-year daily surface solar radiation dataset over China. Science China Earth Sciences, 56, 1555–1565, https://doi.org/10.1007/s11430-012-4542-9, 2013.

Tanner, C. B.: Measurement of Evapotranspiration. In Irrigation of Agricultural Lands. pp. 534–574, https://doi.org/10.2134/agronmonogr11.c30, 1967.

Valiantzas, J. D.: Temperature-and humidity-based simplified Penman's ET0 formulae. Comparisons with temperature-based Hargreaves-Samani and other methodologies. Agricultural Water Management, 208, 326–334, https://doi.org/10.1016/j.agwat.2018.06.028, 2018.

Wanniarachchi, S., & Sarukkalige, R.: A Review on Evapotranspiration Estimation in Agricultural Water Management: Past, Present, and Future. Hydrology, 9(7), 123, https://doi.org/10.3390/hydrology9070123, 2022.

Wu, L., Peng, Y., Fan, J., & Wang, Y.: Machine learning models for the estimation of monthly mean daily reference evapotranspiration based on cross-station and synthetic data. Hydrology Research, 50, 1730–1750, https://doi.org/10.2166/nh.2019.060, 2019.

Yang, K., Koike, T., & Ye, B.: Improving estimation of hourly, daily, and monthly solar radiation by importing global data sets. Agricultural and Forest Meteorology, 137, 43–55, https://doi.org/10.1016/j.agrformet.2006.02.001, 2006.

Yin, J., Deng, Z., Ines, A. V. M., Wu, J., & Rasu, E.: Forecast of short-term daily reference evapotranspiration under limited meteorological variables using a hybrid bi-directional long short-term memory model (Bi-LSTM). Agricultural Water Management, 242, 106386, https://doi.org/10.1016/j.agwat.2020.106386, 2020.

Zhou, N., Wu, L., Yang, Q., Yang, L., Dong, J., & Li, Y.: A high-quality gap-filled daily ETo dataset for China during 1951-2021 from synoptic stations [Data set]. Zenodo. https://doi.org/10.5281/zenodo.11496932, 2024.