DECEMBER 2025
No 395

Authors
Thanh Nguyen-Xuan
Thanh Ngo-Duc
Tung Nguyen-Duy
Tam Nguyen-Ngoc-Minh
Dzung Nguyen-Le
Thi Thanh Nga Pham
Frédéric Thomas

Coordination
Marie-Noëlle Woillez

Research papers

Developing Historical Gridded Climate Datasets for Vietnam Using Reanalysis Data and Station Observations Since the 1940s

ÉDITIONS AFD
AGENCE FRANÇAISE DE DÉVELOPPEMENT

Agence française de développement

Papiers de recherche

Les *Papiers de Recherche de l'AFD* ont pour but de diffuser rapidement les résultats de travaux en cours. Ils s'adressent principalement aux chercheurs, aux étudiants et au monde académique. Ils couvrent l'ensemble des sujets de travail de l'AFD : analyse économique, théorie économique, analyse des politiques publiques, sciences de l'ingénieur, sociologie, géographie et anthropologie. Une publication dans les *Papiers de Recherche de l'AFD* n'en exclut aucune autre.

Les opinions exprimées dans ce papier sont celles de son (ses) auteur(s) et ne reflètent pas nécessairement celles de l'AFD. Ce document est publié sous l'entière responsabilité de son (ses) auteur(s) ou des institutions partenaires.

Research Papers

AFD Research Papers are intended to rapidly disseminate findings of ongoing work and mainly target researchers, students and the wider academic community. They cover the full range of AFD work, including: economic analysis, economic theory, policy analysis, engineering sciences, sociology, geography and anthropology. *AFD Research Papers* and other publications are not mutually exclusive.

The opinions expressed in this paper are those of the author(s) and do not necessarily reflect the position of AFD. It is therefore published under the sole responsibility of its author(s) or its partner institutions.

Developing Historical Gridded Climate Datasets for Vietnam Using Reanalysis Data and Station Observations Since the 1940s

Authors

Thanh Nguyen-Xuan
University of Science and Technology of Hanoi (USTH), Vietnam Academy of Science and Technology (VAST), Hanoi, Vietnam

Thanh Ngo-Duc*
University of Science and Technology of Hanoi (USTH), Vietnam Academy of Science and Technology (VAST), Hanoi, Vietnam
ngo-duc.thanh@usth.edu.vn

Tung Nguyen-Duy
Oxford University Clinical Research Unit, Ho Chi Minh city, Vietnam

Tam Nguyen-Ngoc-Minh
University of Science and Technology of Hanoi (USTH), Vietnam Academy of Science and Technology (VAST), Hanoi, Vietnam

Dzung Nguyen-Le
University of Science and Technology of Hanoi (USTH), Vietnam Academy of Science and Technology (VAST), Hanoi, Vietnam

Thi Thanh Nga Pham
Institute of Meteorology, Hydrology and Climate Change (IMHEN), Ministry of Agriculture and Environment, Hanoi, Vietnam

Frédéric Thomas
Institut de Recherche pour le Développement UMR SENS (IRD, CIRAD, UMPV, Univ Montpellier), Montpellier, France

* Corresponding author

Coordination

Marie-Noëlle Woillez (AFD)

Abstract
Accurate and high-resolution gridded climate data are crucial for scientific applications, yet such data have long faced limitations due to sparse and uneven station networks. In this study, we address this gap for Vietnam by developing the Vietnam Historical Gridded Climate (VnHGC) dataset, which for the first time combines an expanded suite of historical station observations, including newly digitized and quality-controlled records with state-of-the-art reanalysis products through Optimal Interpolation (OI). Two products were constructed: a monthly dataset for 1940–1953 with a horizontal resolution of 25 km based on the background field of ERA5, and a daily dataset for 1961–1980 with a horizontal resolution of 10 km based on the background field of ERA5-LAND. Historical observations were bias-corrected and assimilated into the background fields to maximize local representativeness, especially in data-poor regions. In general, evaluation against independent station data shows that VnHGC substantially reduces systematic biases and better resolves regional patterns of temperature and precipitation, with the greatest improvements in northern and mountainous areas where reanalyses typically underperform. These improvements are especially pronounced in the daily dataset (1961–1980), reflecting the combined effects of higher temporal resolution and greater station availability.

Keywords
gridded dataset, climate data, Vietnam

Acknowledgements
This research was supported as part of the second phase of the GEMMES Vietnam project, funded by the French Development Agency (AFD) through the Facility 2050. We would like to express our special thanks to Tuyet Nguyen and Tra Nguyen, members of the REMOSAT research group at USTH, and Nam Nguyen (IMHEN), for their contribution to the digitization and preparation of some of the data used in this article. We also thank Dr. Marie-Noëlle Woillez (AFD) for her careful review and valuable suggestions that helped improve this manuscript.

Original version
English

Accepted
December 2025

Résumé

Des données climatiques maillées précises et à haute résolution sont essentielles pour de nombreuses applications scientifiques, mais elles ont longtemps été limitées par la rareté et la répartition inégale des stations d'observation. Dans cette étude, nous comblons cette lacune pour le Vietnam en développant le Vietnam Historical Gridded Climate (VnHGC) dataset, qui combine pour la première fois un ensemble élargi d'observations historiques provenant de stations, incluant des données d'archives récemment numérisées et contrôlées en qualité, avec des produits de réanalyse de pointe grâce à l'Interpolation Optimale (OI). Deux produits ont été élaborés : un jeu de données mensuelles couvrant la période 1940–1953 avec une résolution horizontale de 25 km, basé sur le champ de fond d'ERA5, et un jeu de données journalières couvrant 1961–1980 avec une résolution horizontale de 10 km, basé sur le champ de fond d'ERA5-LAND. Les observations historiques ont été corrigées des biais et assimilées dans les champs de fond afin de maximiser leur représentativité locale, notamment dans les régions pauvres en données. De manière générale, l'évaluation par rapport à des stations indépendantes montre que VnHGC réduit considérablement les biais systématiques et représente mieux les structures régionales de la température et des précipitations, avec des améliorations particulièrement marquées dans les régions septentrionales et montagneuses où les réanalyses présentent généralement des performances moindres. Ces améliorations sont encore plus nettes pour le jeu de données journalières (1961–1980), ce qui reflète les effets combinés d'une résolution temporelle plus fine et d'une disponibilité accrue des stations.

Mots-clés

jeu de données sur grille, données climatiques, Vietnam

Remerciements

Cette recherche a été soutenue dans le cadre de la deuxième phase du projet GEMMES Vietnam, financé par l'Agence Française de Développement à travers la Facilité2050. Nous souhaitons exprimer nos remerciements particuliers à Tuyet Nguyen et Tra Nguyen, membres du groupe de recherche REMOSAT à l'USTH, ainsi qu'à Nam Nguyen (IMHEN), pour leur contribution à la numérisation et à la préparation de certaines données exploitées dans cet article. Nous remercions également le Dr Marie-Noëlle Woillez (AFD) pour sa relecture attentive et ses précieuses suggestions qui ont permis d'améliorer ce manuscrit.

Version originale

Anglais

Acceptée

Décembre 2025

Introduction

Accurate and high-resolution meteorological data are essential for a wide range of scientific studies. Although station-based observations are regarded as the most reliable, their sparse and uneven distribution limits their applicability for fine-scale analyses. Therefore, the development of long-term, high-resolution, and consistent gridded datasets for meteorological variables, especially for rainfall and temperature, is of significant importance.

Gridded datasets offer several advantages. They provide valuable climate information in regions with limited data availability and ensure spatial uniformity, simplifying spatial averaging and large-scale analyses. They can also serve as inputs for various applications, such as extreme events analysis (Bhattacharyya et al., 2021), hydrological modeling (Hong et al., 2007), assessment of crop response to weather variability (Parkes et al., 2019), and validation of climate models (Trinh et al., 2021). Additionally, they are well-suited for studies employing pattern recognition techniques like Empirical Orthogonal Functions (EOF). For instance, using the Vietnam Gridded Precipitation (VnGP) dataset (Nguyen-Xuan et al., 2016), Tuan (2019) applied EOF analysis to investigate rainfall patterns in Vietnam, revealing that intra-seasonal rainfall variations differ across regions and are strongly influenced by local topography.

Despite the critical need for high-resolution meteorological data, the limitations posed by observational gaps and limited resources pose a significant challenge in producing globally consistent and detailed datasets. A well-known source for global gridded datasets is the reanalysis products from the European Centre for Medium-Range Weather Forecasts (ECMWF). These products include the fifth-generation atmospheric reanalysis (ERA5) with a spatial resolution of 0.25° (~25 km; Hersbach et al., 2020) and its land-focused replay with a finer spatial resolution of 0.1°, ERA5-Land (~10 km; Muñoz-Sabater et al., 2021). Both datasets are widely used for various research topics, but their resolution and accuracy may still be insufficient for local-scale studies. Therefore, many efforts have been made to develop regionally high-resolution datasets (Aybar et al., 2019; Serrano-Notivoli et al., 2017; Lussana et al., 2018). These efforts encompass a range of objectives and methodologies. For example, Aybar et al. (2019) created a gridded rainfall dataset for Peru at a resolution of 0.1° spanning 1981 to the present. Serrano-Notivoli et al. (2017) developed the Spanish PREcipitation At Daily scale (SPREAD) dataset, a 5 km daily precipitation product for Spain based on over 12,800 stations on the mainland and islands. Even higher-resolution datasets exist, such as seNorge2, which provides 1 km daily precipitation fields for mainland Norway from 1957 onward, offering valuable inputs for climate and hydrological studies in complex terrain (Lussana et al., 2018).

In Vietnam, there are still relatively few gridded datasets that are specifically constructed for applications from regional to local scales. Certain studies used data from ERA5 and ERA5-LAND, but the accuracy of these datasets in Vietnam remains a concern. For instance, Roversi et al. (2024) reported that ERA5-LAND struggled to capture extreme rainfall events in 2020 in Vietnam. Nguyen-Xuan et al. (2025) conducted an analysis for Hanoi and Ho Chi Minh City, revealing that ERA5 exhibited poorer temperature patterns compared to an observation-based gridded product, the daily Vietnam Gridded Climate Dataset (VnGC) (Tran-Anh et al., 2023). Nguyen-Xuan et al. (2025) also found that both ERA5 and VnGC failed to capture stational patterns when compared to stational observations. Cold biases were observed in high-density cities, while warm biases were found in mountainous areas, with the biases in ERA5 being more pronounced. The biases were partly attributed to the coarse spatial resolutions of VnGC (10 km) and ERA5 (25 km), which do not allow these datasets to fully capture urban heat island effects, as simulations conducted at higher resolution (2km) showed improved performance. It is worth noting that the VnGC is created at a higher resolution than ERA5 and is based on a denser density of station observations. The biases observed in VnGC may also be attributed to the interpolation method used to create the data, which could have potentially smoothed out signals from certain stations in specific areas.

Satellite-based observations, in addition to reanalysis data, also serve as an important source of meteorological information. They are particularly useful in locations where rain gauges are scarce, as suggested by Vu et al. (2012), Ngo-Duc et al. (2013), and Hiep et al. (2018). However, the accuracy of satellite data can vary significantly, especially in areas with complex topography. Consequently, applying bias correction is recommended to enhance the reliability of satellite outputs (Ngo-Duc et al, 2013; Nguyen et al., 2024; Le et al., 2024). Other datasets, such as CHIRPS (Funk et al., 2015) and GSMaP (Kubota et al., 2020), are developed by blending satellite and gauge-based data. However, these datasets rely on a limited number of rain gauges in Vietnam, potentially leading to higher uncertainty, particularly in regions with few or no in-situ observations.

To address the scarcity of reliable gridded rainfall datasets for Vietnam, several studies have been conducted to develop such datasets. Nguyen-Xuan et al. (2016) developed the first high-resolution (10-km) daily dataset of rainfall, utilizing nearly 500 rain gauges and spanning the period from 1980 to 2010. This dataset, known as the Vietnam Gridded Precipitation dataset (VnGP), has gained significant recognition and has been widely used in various studies. For instance, it has been employed in validating the performance of climate and hydrological models (Trinh et al., 2021; Tran et al., 2025), assessing the skill of precipitation forecasts (Main et al., 2024), validating satellite-based rainfall estimates (Nodzu et al., 2019; Trinh-Tuan et al., 2019b), exploring the relationship between large-scale climate patterns and local rainfall variability (Bui-Minh et al., 2024), investigating atmospheric mechanisms associated with rainfall changes in central Vietnam (Van et al., 2023), and studying the onset of the second rainy season in the Central Highlands (Bui-Manh et al., 2021). Beyond its direct applications, the VnGP dataset serves as a valuable reference for constructing transfer functions for correcting bias in future rainfall projections (Trinh-Tuan et al., 2019a). Its creation methodology has been applied in the development of the aforementioned VnGC dataset (Tran-Anh et al., 2023).

In this study, we present a novel gridded dataset for Vietnam, utilizing newly collected data from multiple sources, including historical station records spanning from 1940 to 1980. We employed a data assimilation method to adapt to the features of the newly collected data. This dataset, which covers daily and monthly rainfall and temperature data, extends the coverage period beyond the VnGP (1980–2010) and VnGC (1980–2019), enabling researchers to conduct studies spanning extended historical periods. Specifically, we have constructed two datasets, including: 1) a daily dataset covering 1961 to 1980; and 2) a monthly dataset covering the period from 1940 to 1953. These new datasets, named the Vietnam Historical Gridded Climate (VnHGC) dataset, offer valuable insights into past temperature and precipitation patterns and their potential connections to natural and societal conditions. It is worth noting that a parallel effort to develop a more recent gridded dataset (1980–present) is currently underway in collaboration with the Institute of Meteorology, Hydrology and Climate Change (IMHEN), based on a much denser station network. Therefore, the present work, which focuses on the period prior to 1980, serves as a complementary contribution.

The paper is organized as follows. Section 1 describes the data sources and methodology. Section 2 presents the performance evaluation, and Section 3 concludes with a concise summary.

1. Data and Methodology

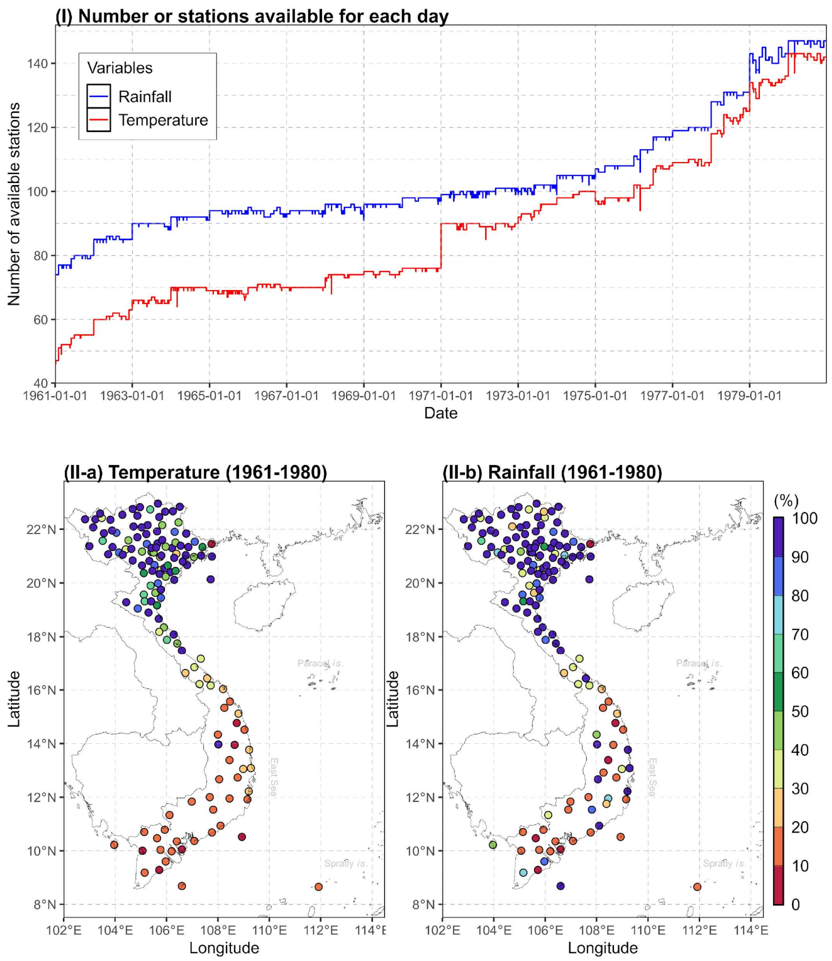
1.1. Station data

In this study, we use two sets of station data. The first dataset, referred to as OBS-1, includes daily temperature and rainfall observations from up to 147 stations spanning the period from 1961 to 1980. The data were collected from the Vietnam Meteorological Hydrological Administration (VNMHA) and pre-processed by IMHEN. Figure 1 presents a comprehensive analysis of the data availability statistics. It indicates a gradual increase in the number of available stations over time, with a notable rise occurring after 1975. Notably, the number of temperature stations was consistently lower compared to the number of rainfall stations. Furthermore, Figure 1 reveals that stations in North Vietnam provided data for a substantial portion of the time, ranging from 60% to over 90%, during the period from 1961 to 1980. This coverage is significantly higher than in the South, where time coverage was generally below 30% due to the historical impacts of the Vietnam war (1954-1975).

The second dataset, referred to as OBS-2, comprises monthly temperature and rainfall observations collected and digitized from historical archives spanning the period 1922–1953. Derived from the Annuaire Statistique de l'Indochine (*ASI*; Thomas & Nguyen, 2024a), this dataset represents a novel and unique historical resource that has not previously been used for analyzing Vietnamese climatology. The *ASI* contains annual synthesis reports produced by the French administration, with the first chapter dedicated to climatological information. Similar to Figure 1, Figure 2 provides statistics on data availability (i.e., data collected and already digitized by our team) over the period 1922–1953, highlighting notable gaps caused not by a limited network of stations — there was already an extensive network of weather and rainfall stations in place during this period — but rather by a lack of resources to publish monthly weather data from each station on an annual basis. For example, the first volume of the *ASI*, published in 1927, chose to publish the average monthly temperature and precipitation data from 17 weather stations for the period from 1906 to 1922 (rather than monthly data for each year). However, this volume did publish monthly data from the same stations for the year 1922. Volume 2, published in 1931, which was supposed to cover the period from 1923 to 1929, made the editorial choice to publish the average monthly temperature and precipitation data from 53 stations for periods that varied depending on the station. It was not until the third volume that monthly data were published more systematically on a year-by-year basis. These editorial choices, which were justified by the desire to establish monthly averages for temperature and precipitation over long periods in order to better understand the climates of the Indochina Peninsula, explain the numerous gaps in Figure 2 for the 1920s. Additionally, as illustrated in Figure 2, the archive includes data from Cambodia, Laos, and Vietnam, which may slightly dilute the focus on Vietnam.

These discontinuities are expected to be resolved through the ongoing integration of additional digitized sources, such as the *Bulletin mensuel des observations* (Thomas and Nguyen, 2024b). Notably, the historical station dataset used here was recovered from archival holdings and is more thoroughly examined in Thomas et al. (2025). The data are now accessible online[1], albeit currently under embargo, and will ultimately be released as open access, thereby facilitating broader scientific use. As we continue to digitize and integrate additional historical datasets, we anticipate expanding these gridded climatological datasets for Vietnam in the future.

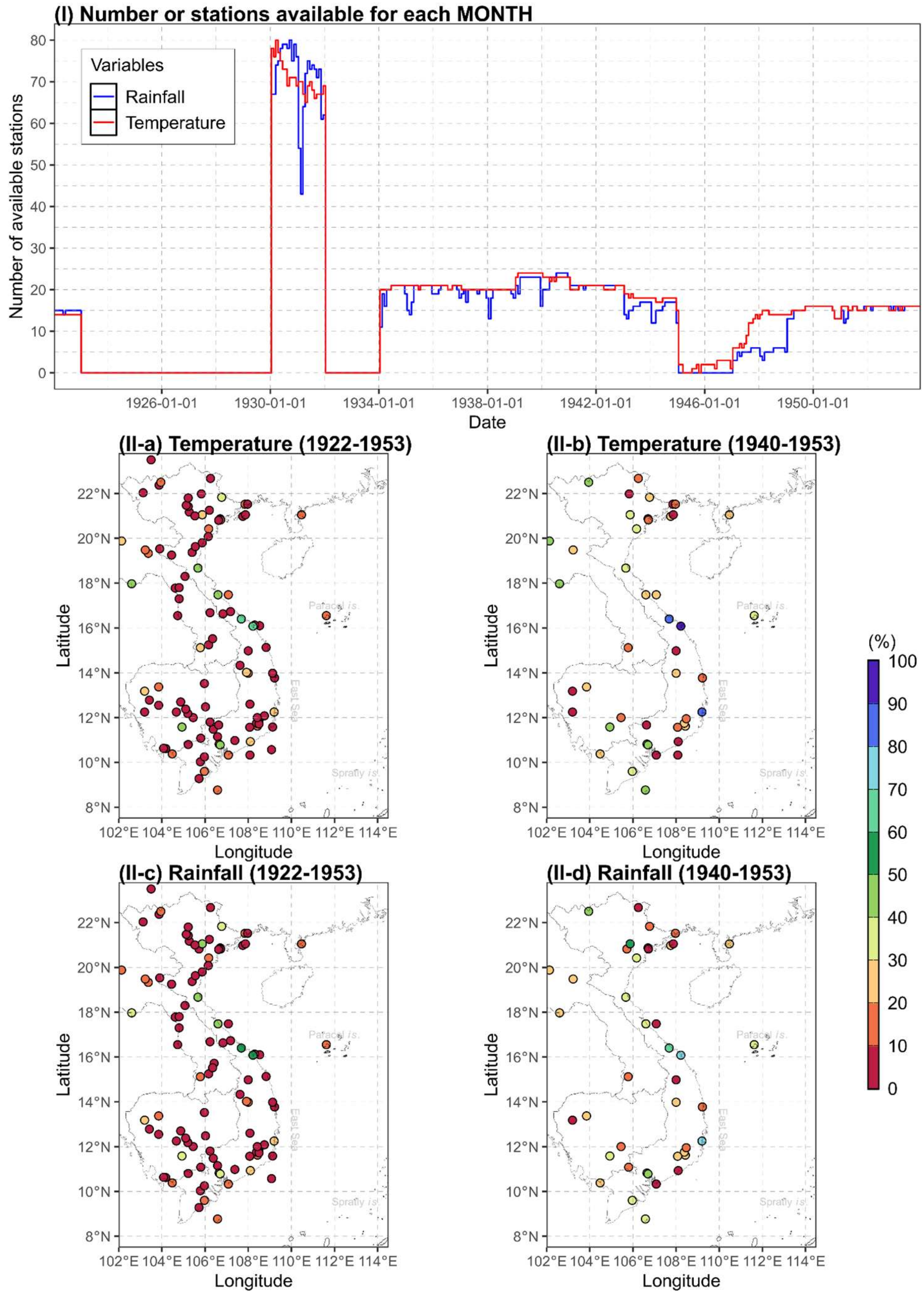
Figure 1. Statistics of OBS-1's data availability.



Note: (I) number of available stations per day (unit: stations/day) during the period 1961–1980; and (II) percentage of usable data (unit: %) at each station during 1961–1980, estimated separately for temperature (II-a) and rainfall (II-b). Source: data collected from the VNMHA. Authors' own visualization, original.

[1]https://dataverse.ird.fr/dataverse/climate_vietnam_cambodia_laos

Figure 2. Statistics of OBS-2's data availability.



Note: (I) the number of stations with data already digitized at each month (unit: stations/month) during the period 1922–1953; and (II) the percentage of usable data (unit: %) at each station during the period 1922–1953 (left sub-panels) and 1940–1953 (right sub-panels) estimated separately for temperature (top sub-panels) and rainfall (bottom sub-panels). Source: data collected from the *ASI*. Authors' own visualization, original.

1.2. ERA5 and ERA5-Land

Traditional interpolation methods, such as Spheremapping used in VnGC and VnGP, typically yield more accurate results when a larger number of stations are used in the interpolation process. However, this approach may introduce inconsistencies in the results if the number of available stations varies over time or if the spatial distribution of stations is unevenly distributed and differs between time steps. As discussed in the previous section, our stationary datasets, OBS-1 and OBS-2, exhibit distinct characteristics, and both datasets experience a notable temporal fluctuation in the availability of stations. Therefore, the conventional interpolation method may not be suitable for our case. Instead, the data assimilation technique, which essentially employs station data to enhance a pre-prepared background field, emerges as a more appropriate approach. A detailed explanation of this technique will be provided in the subsequent section, but here, we briefly introduce the data used as the background field.

For use with OBS-1, spanning from 1961 to 1980, the ERA5-Land dataset (Muñoz-Sabater et al., 2021) was chosen. Developed by the ECMWF under the Copernicus Climate Change Service (C3S), ERA5-Land is an enhanced global land dataset that provides hourly data at a high horizontal resolution of approximately 10 kilometers. This dataset covers the period from 1950 to the present. For this study, ERA5-Land data were obtained from the Copernicus Climate Data Store (CCDS) at an hourly resolution and adjusted to Local Solar Time (GMT+7). Subsequently, daily accumulated rainfall and mean temperature were calculated over a 24-hour period, starting from 19:00 PM of one day and ending at 19:00 PM of the next, in accordance with the real-world observation practice in Vietnam. The grid of the 10-kilometer resolution of ERA5-Land was retained in the gridded datasets constructed for this period.

For OBS-2, spanning back to 1922, data availability constraints prevented the use of ERA5-Land, which is only available from 1950 onward. Instead, ERA5 monthly reanalysis data were utilized. Developed by the ECMWF within the same C3S framework, ERA5 provides global coverage from 1940 onward and is available at both hourly and monthly time scales (Hersbach et al., 2019). While its horizontal resolution (~25 km) is coarser than that of ERA5-Land, ERA5 covers a longer historical period and includes both surface and pressure-level variables. In this study, we also use the monthly mean data obtained from the CCDS. Although OBS-2 extends back to 1922, the period covered by ERA5 (from 1940 onward) limits the temporal coverage of the new VnHGC dataset built in this study to start in 1940.

1.3. Methodology

We employ the widely known data assimilation technique, Optimal Interpolation (OI), as described by Kalnay (2003), with a few slight modifications to accommodate the unique characteristics of our dataset. OI is particularly advantageous because it integrates observations with a priori estimates, known as the background, to produce a more accurate estimate of a field by minimizing the error variance. As illustrated in Figures 1 and 2, the availability of our station datasets, particularly in OBS-2 data, is limited over the two study periods and exhibits spatial inconsistency across the country. Consequently, rather than utilizing traditional interpolation methods, such as those employed in VnGC and VnGP, which heavily rely on the quantity and density of station observations, the OI algorithm, which primarily utilizes background data and enhances it with observations, is more appropriate for this study.

OI has been extensively used due to its simplicity and effectiveness, as evidenced by previous studies in the field of data assimilation (e.g., Huang et al., 2020). Figure 3 provides a schematic diagram illustrating the overall computing process and primary equations employed in the OI technique. The main computation processes of OI are succinctly described in the following 8 steps:

1. **Input preparation**: The input data include the background field from reanalysis products (i.e., ERA5-LAND or ERA5), denoted as $\mathbf{x}^b$, and the raw observations at stations, denoted as $\mathbf{y}^o_{raw}$.

2. **Perform bias correction on station observations:** Observations in various fields often show sparsity, localization, and site-specific errors like sensor malfunctions. These biases are easier to identify and correct at the station level. Meanwhile, background data from reanalysis is spatially consistent and adjusted over time. Thus, simple bias correction of observations is essential to ensure discrepancies with the background are due to genuine variability rather than systematic errors. This enhances the reliability of interpolation processes. The bias corrected observations are obtained as:

$$\mathbf{y}^o = \mathbf{y}^o_{raw} - mean\left(\mathbf{y}^o_{raw} - \mathbf{H}(\mathbf{x}^b)\right) \quad (1)$$

where $\mathbf{H}(\mathbf{x}^b)$ is the "first guess" of the observations, i.e., the background is interpolated to the observation location with $\mathbf{H}$ being the operator that performs the necessary interpolation and transformation from $\mathbf{x}^b$ to observation space.

3. **Estimate correlation length:** The correlation length ($\boldsymbol{L}$) is a key parameter that determines the extent to which observations impact the background. Due to the significant temporal variation in the number of stations, which results in fluctuations in station density for interpolation, a cross-validation-based optimization is conducted at each time step. This process involves estimating $\boldsymbol{L}$ by fitting an empirical semivariogram using pairwise distances and squared differences.

4. **Estimate background and observation error variances:** These variances are used to compute background (ϵ_b^2) and observation (ϵ_o^2) error covariance matrices, which are essential in determining how much weight is given to the background versus the observations when computing the OI analysis, $\mathbf{x}^a$, which is the best estimate of the true state. Although these variances are important in determining how much confidence you have in the background field compared to observations, estimating these variances is usually one of the most critical and challenging tasks in OI. For simplicity and to save computing resources, Bayesian estimation is used to infer background and observation error variances by analyzing how well different variance values explain the mismatch between observations and background. By minimizing cross-validation error or maximizing likelihood, we estimate the most probable variances, ensuring an optimal balance between model and observations based on their uncertainties.

5. **Compute background and observation error covariance matrices:** The background ($\mathbf{B}$) and observation ($\mathbf{R}$) error covariance matrices are then computed as follows:

$$\mathbf{B} = \epsilon_b^2 . exp(-\boldsymbol{d}^2 / \boldsymbol{L}^2) \quad (2)$$

$$\mathbf{R} = diag(\epsilon_o^2) \quad (3)$$

6. **Compute Weights and Innovations:** The weights, **W**, are a function of the distance between the observation and the grid point, and the analysis is iterated several times. In OI (Gandin, 1963), the matrix of weights **W** is determined from the minimization of the analysis errors at each grid point. The optimal weight matrix is used to minimize the analysis covariance:

$$\mathbf{W} = \mathbf{BH}^T(\mathbf{R} + \mathbf{HBH}^T)^{-1} \quad (4)$$

The observational increments, or "innovations", are calculated by subtracting the model's first guess, $\mathbf{H}(\mathbf{x}^b)$, from the observation ($\mathbf{y}^o$), resulting in:

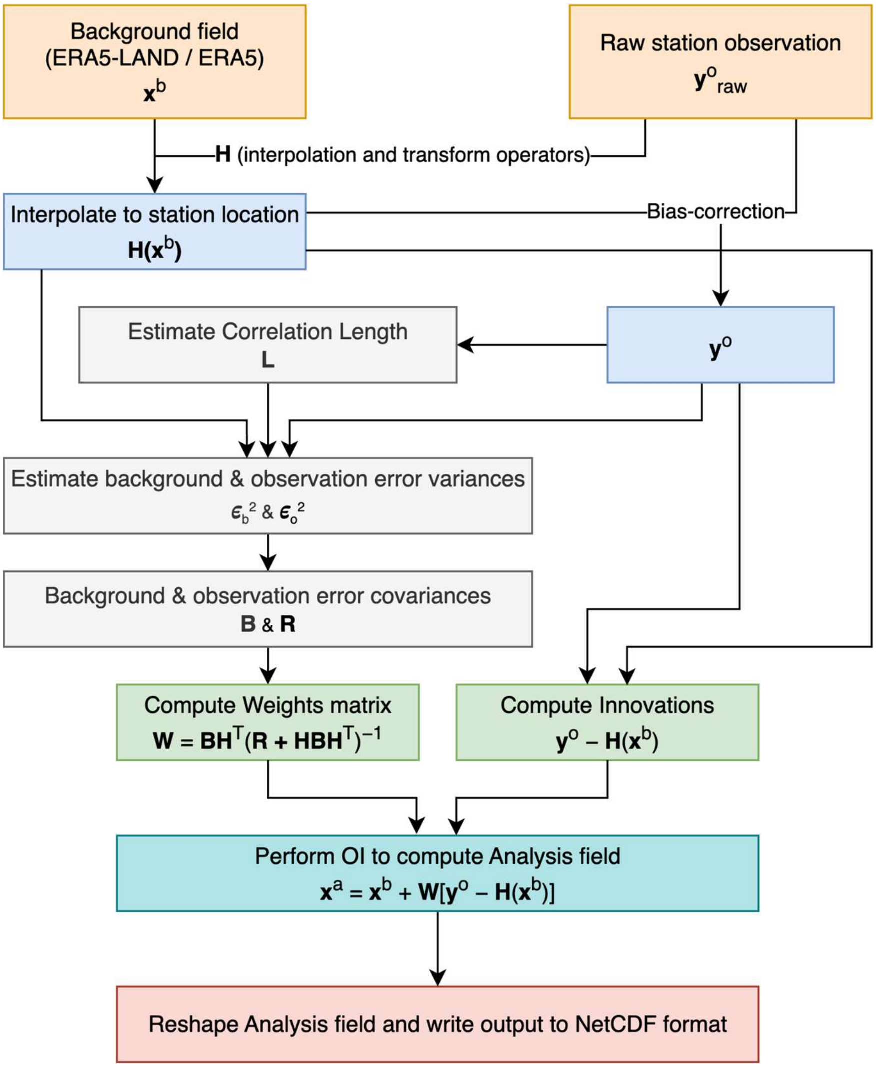
$$\mathbf{y}^o - \mathbf{H}(\mathbf{x}^b) \quad (5)$$

7. **Perform OI to compute analysis field**: the OI analysis, $\mathbf{x}^a$, is determined by adding the innovations to the first guess, with weights (denoted as $\mathbf{W}$) that are based on the estimated statistical error covariances of the background field and station observations:

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{W}[\mathbf{y}^o - \mathbf{H}(\mathbf{x}^b)] \qquad (6)$$

The analysis field is then reshaped and written into NetCDF format.

Figure 3. Schematic diagram illustrating the process for constructing the gridded climate dataset through the Optimal Interpolation (OI) method.



Source: Authors' own visualization. Original.

2. Results

2.1. Evaluation for the period 1961–1980

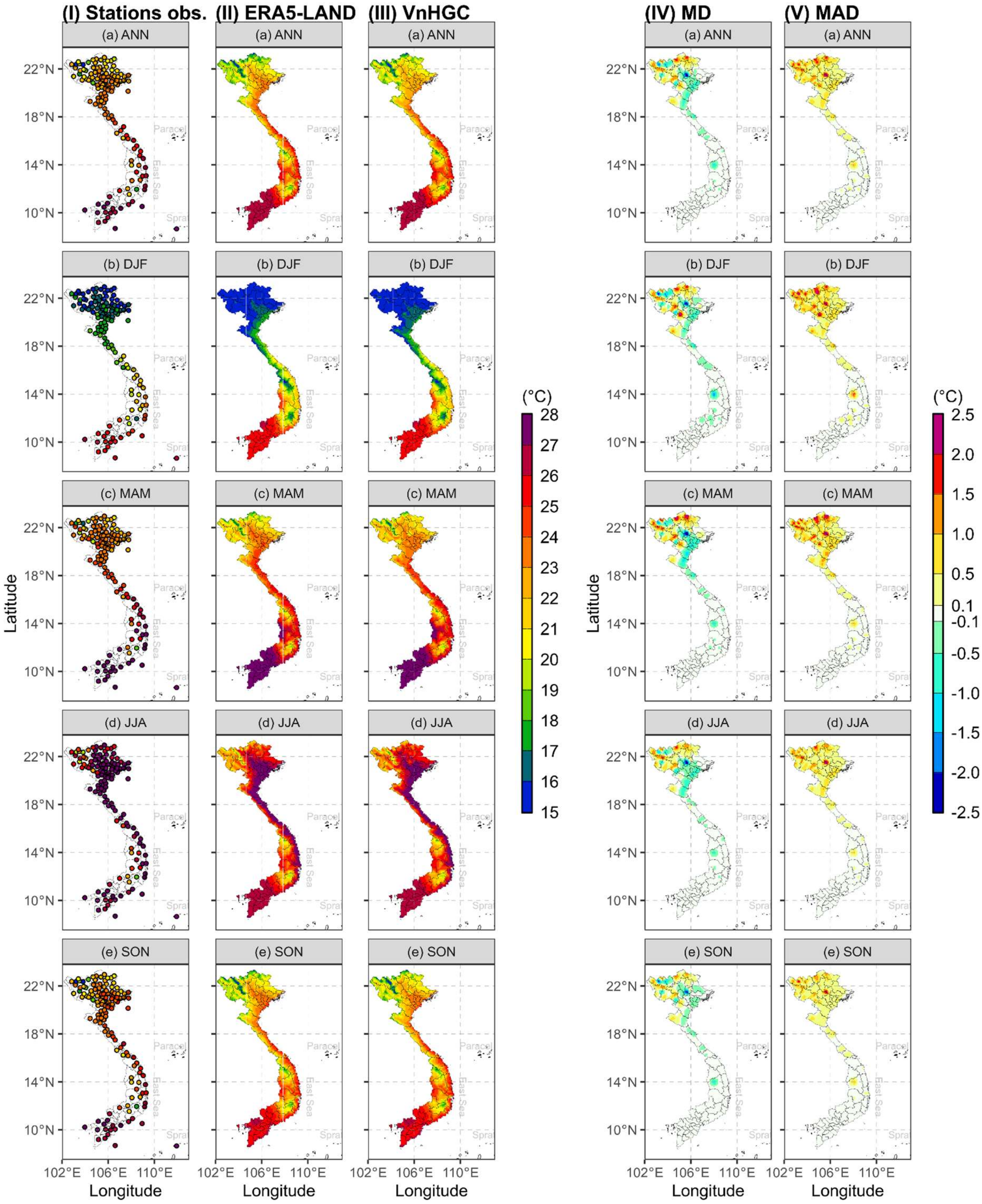
We commenced the quality assessment of the dataset by comparing the climatological means over the entire period and the seasonal climatological means between VnHGC, its driving background fields from ERA5-Land, and raw station observations. It's important to note that for the 1961–1980 period, the VnHGC and ERA5-Land datasets were gridded at a daily temporal resolution with a horizontal resolution of 10-km.

The results presented in Figures 4 I–III and 5 I–III reveal a striking similarity in climatological means across the datasets. Notably, the ERA5-LAND reanalysis consistently captures the spatial distribution of temperature and precipitation effectively during the study period. When comparing panels II and III in both Figures 4 and 5, the differences between ERA5-LAND and the newly developed VnHGC dataset are barely discernible by visual inspection. Moreover, both datasets exhibit a remarkably high level of agreement with station observations.

The differences between the new VnHGC dataset and ERA5-LAND become evident when analyzing panels IV and V, which illustrate the mean difference (MD) and the mean absolute difference (MAD). In general, when comparing the MD and MAD values, the results for precipitation and temperature reveal distinct patterns.

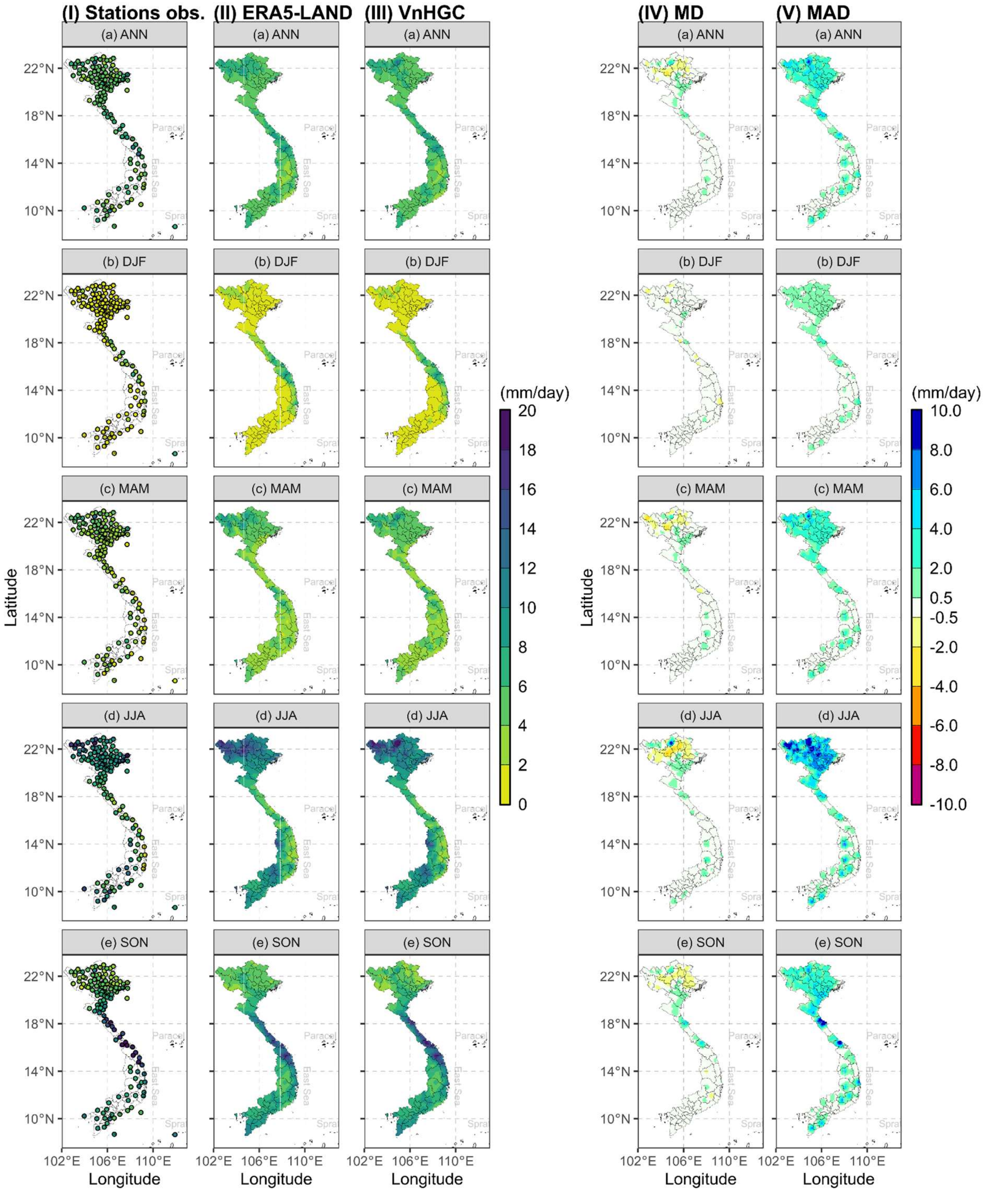
Temperature differences are most pronounced in northern regions, with mean variations reaching up to ±1.5 °C. In contrast, areas south of latitude 17° experience negligible temperature differences, except for localized zones in the Central Highlands. This disparity can be attributed to the limited station observations in southern regions, as discussed in relation to Figure 1. Consequently, the OI algorithm was largely inactive in these regions, resulting in ERA5-LAND results that remained largely unchanged. A similar situation is observed for daily precipitation. Furthermore, the magnitude of temperature differences between the two datasets is relatively consistent across both climatological and seasonal means. ERA5-LAND tends to overestimate temperature values in lowland areas and underestimate them in mountainous regions.

Figure 4. Spatial distribution of daily mean temperature (°C) averaged for 1961-1980.



Note: (a) over the entire period, and (b-e) for each season (DJF, MAM, JJA, and SON). Data are estimated from (I) raw station observations OBS-1, computed only over the periods within 1961–1980 for which data are available, rather than from continuous series, (II) ERA5-LAND, and (III) new gridded VnHGC. Panels (IV) and (V) show the mean difference (MD) and mean absolute difference (MAD) between VnHGC and ERA5-LAND, respectively. Source: Authors' own calculation, original.

Figure 5. Similar to Figure 4, but for daily rainfall (mm/day).



Source: Authors' own calculation, original.

Precipitation data from Figures 5 IV and 5 V reveal significant differences between the two datasets, particularly in northern Vietnam, central Vietnam, and parts of the Central Highlands. These differences are most pronounced during the JJA months, which coincide with the peak rainfall season in these regions. In certain northern areas, particularly high mountain regions, the 20-year JJA mean absolute difference (MAD) values can exceed 10 mm/day. This suggests that ERA5-LAND either failed to capture or only weakly represented localized rainfall effects, which can only be accurately detected through station observations in such areas. VnHGC, on the other hand, incorporates station data that ERA5-LAND may lack or be unable to represent at a 10 km resolution during the reanalysis process, thereby addressing this limitation.

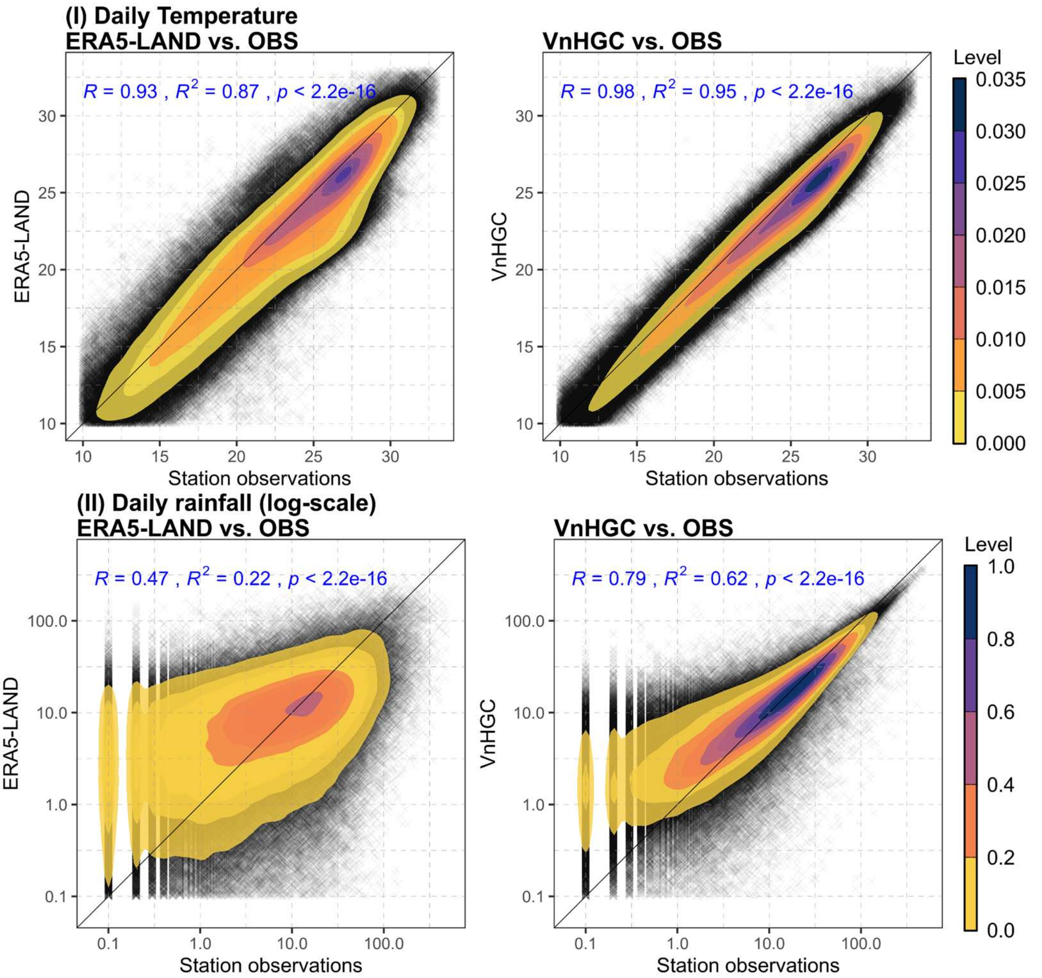
To further clarify the impact of OI on improving ERA5-LAND using station observations, both gridded datasets were interpolated to station locations using nearest-neighbor interpolation. In essence, VnHGC is almost certain to outperform ERA5-LAND when compared with station data, as this is the very purpose of the technique. This is also generally true for gridded datasets constructed using standard interpolation methods, as demonstrated in previous studies on VnGP (Nguyen-Xuan et al., 2016) and VnGC (Tran-Anh et al., 2023). Therefore, the analysis in this section not only reaffirms this point but also examines which aspects exhibit the most pronounced improvements in the new dataset.

Figure 6 presents a comparative analysis of daily precipitation data for nearly 150 stations over the 20-year period (1961–1980), resulting in over one million data points per panel. To enhance clarity, a shaded background is employed to indicate the density of points within different regions of the figure. Higher densities along the main diagonal signify stronger agreement with observations. Additionally, each panel includes several simple statistics, including R, R^2, and statistical significance (p-value). Given the substantial dataset size, the very low p-values are not surprising. Furthermore, since VnHGC incorporates station observations to adjust ERA5-LAND, it is expected and confirmed by the statistics that VnHGC performs better than ERA5-LAND. These results align with the theoretical foundation of the algorithm.

A closer examination of improvements across different regions and for both temperature and precipitation reveals several notable patterns. For temperature, improvements are relatively uniform across the domain. ERA5-LAND already shows high overall agreement with station observations, with most high-density regions located near the main diagonal. However, considerable scatter is also present. Overall, ERA5-LAND tends to fall below the diagonal, indicating a general cold bias relative to observations. Notably, in Figure 6-I, many points deviate significantly from the diagonal, suggesting that ERA5-LAND underestimates temperatures by 1–3 °C in most cases and by up to ~10 °C in some extreme cases. These biases are corrected in VnHGC, demonstrating the effectiveness of incorporating station data to adjust the reanalysis fields.

Improvements in VnHGC are also evident for precipitation, with values clustering more closely around the diagonal compared to ERA5-LAND. It's important to note that precipitation results were transformed into a log-scale to better illustrate the effectiveness of OI in VnHGC. However, since both axes in Figure 6-II are log-scaled, caution is advised when interpreting the figure. A small improvement around 100 mm/day actually represents a much larger correction than relatively large changes in the range of 0.1–1 mm/day, or even 10 mm/day. Additionally, at low precipitation values, ERA5-LAND generally shows a slight wet bias compared to station data. This likely reflects the nature of ERA5-LAND, which incorporates numerical simulations and tends to overproduce non-zero rainfall days—a common feature observed in many modeling studies.

Figure 6. Scatter-plots comparing daily (I) temperature (°C) and (II) rainfall (mm/day) observations with nearest-neighbor interpolated data from ERA5-LAND (left panels) and VnHGC (right panels) during the period 1961-1980.



Note: The shaded area indicates the density of points. For rainfall results, x- and y-axis are log-transformed for better visualization. The shaded area (level) represents the density level of points within each area, with a higher level indicating a higher number of points occupying the same space. Source: Authors' own calculation, original.

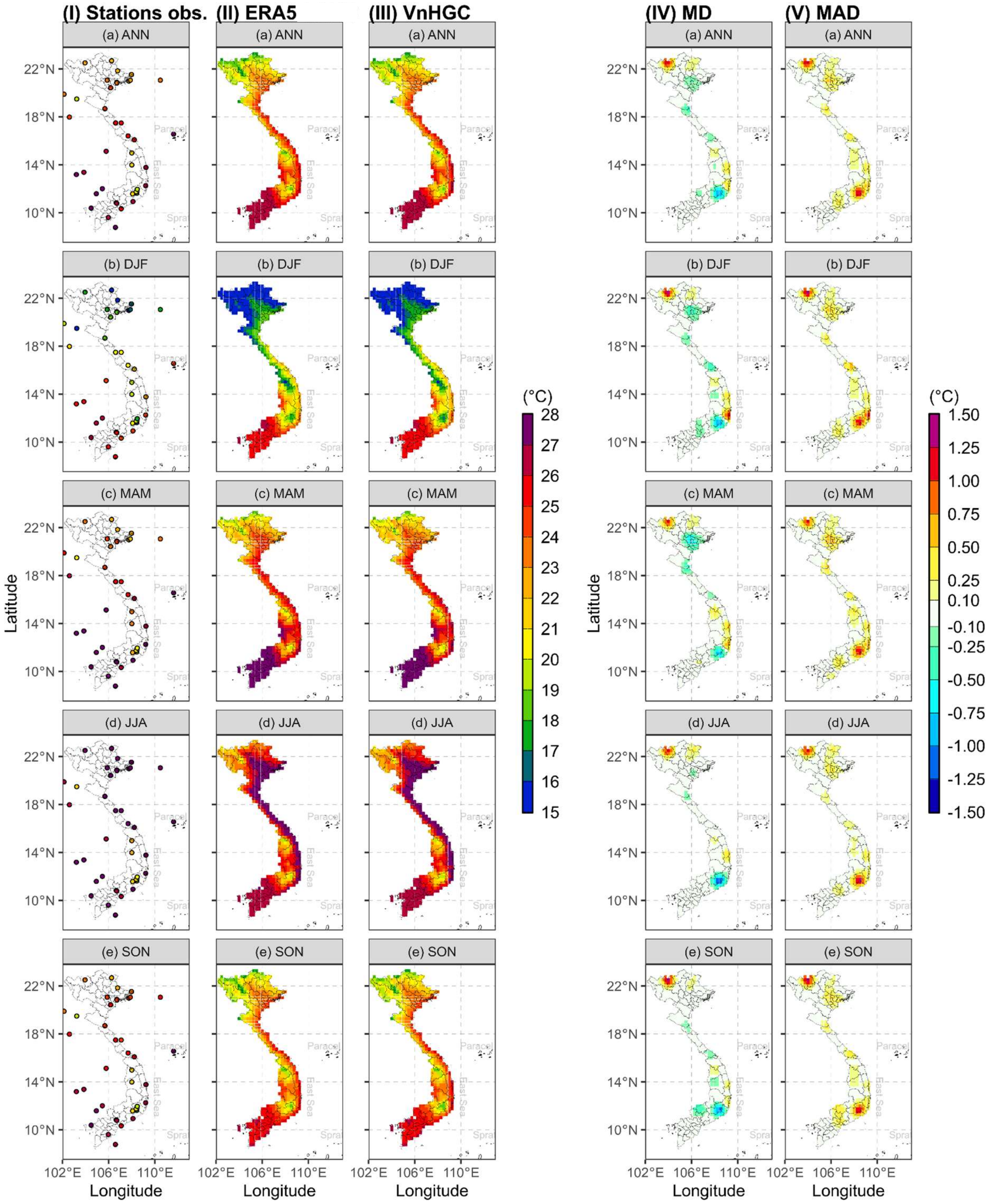
2.2. Evaluation for the period 1940–1953

Similar to the previous section, the same set of analyses was conducted on the gridded dataset spanning from 1940 to 1953. As mentioned earlier, the limited availability of station observations necessitated the construction of this dataset at a monthly temporal scale with a horizontal resolution of 25 km.

Figure 7 compares the 14-year climatological and seasonal means. For both precipitation and temperature, the results are broadly consistent with those of 1960–1981, as discussed in the previous section. ERA5 and VnHGC show close agreement and appear comparable to station observations, with hardly discernible differences by visual inspection. However, when examining the distributions of MD and MAD more closely, the differences between the two datasets become clearer in certain regions. While only a few scattered areas show distinct and large deviations, these differences are noticeable.

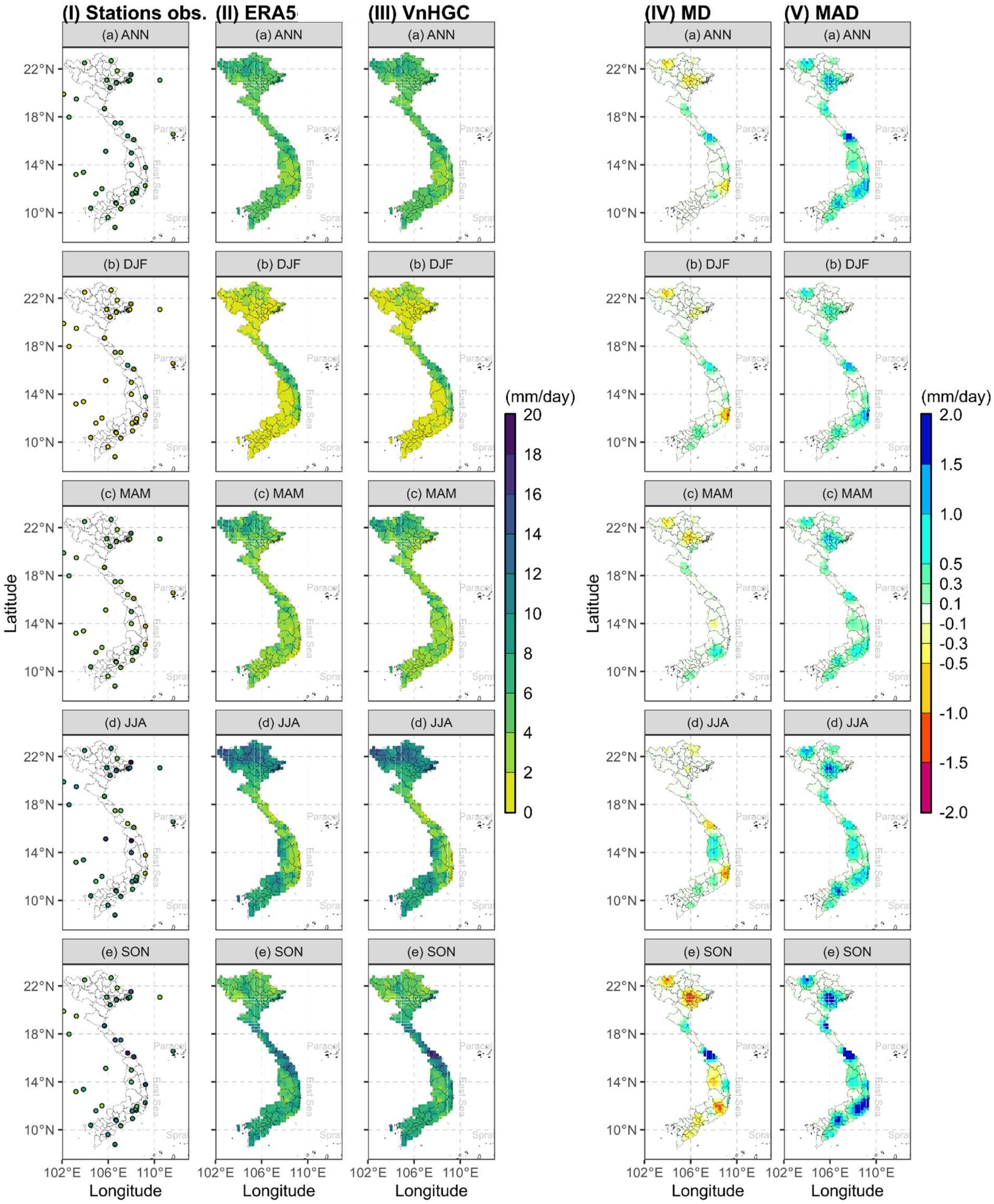
It's important to note that in the original investigation of the OI process, stations located offshore were excluded not only because they are often far from the mainland, but also because their climatic and physical characteristics differ significantly from those on land, which could introduce noise into the calculations. Additionally, for the dataset spanning from 1940 to 1953, we also collected and incorporated station observations from Laos and Cambodia. Nevertheless, as illustrated in Figure II-2b, the distances between these stations and Vietnamese territory are substantial, with only a few stations situated in southern regions. Consequently, the impact of most stations outside Vietnam on the data within Vietnam is negligible.

Figure 7. Spatial distribution of monthly mean temperature (°C) averaged for 1940-1953.



Note: (a) over the entire period, and (b-e) for each season (DJF, MAM, JJA, and SON). Data are estimated from (I) station observations OBS-2, computed only over the periods within 1940–1953 for which data are available, rather than from continuous series, (II) ERA5, and (III) new gridded VnHGC. Panels (IV) and (V) show the mean difference (MD) and mean absolute difference (MAD) between VnHGC and ERA5, respectively.Source: Authors' own calculation, original.

Figure 8. Similar to Figure 7, but for monthly mean of daily rainfall (mm/day).

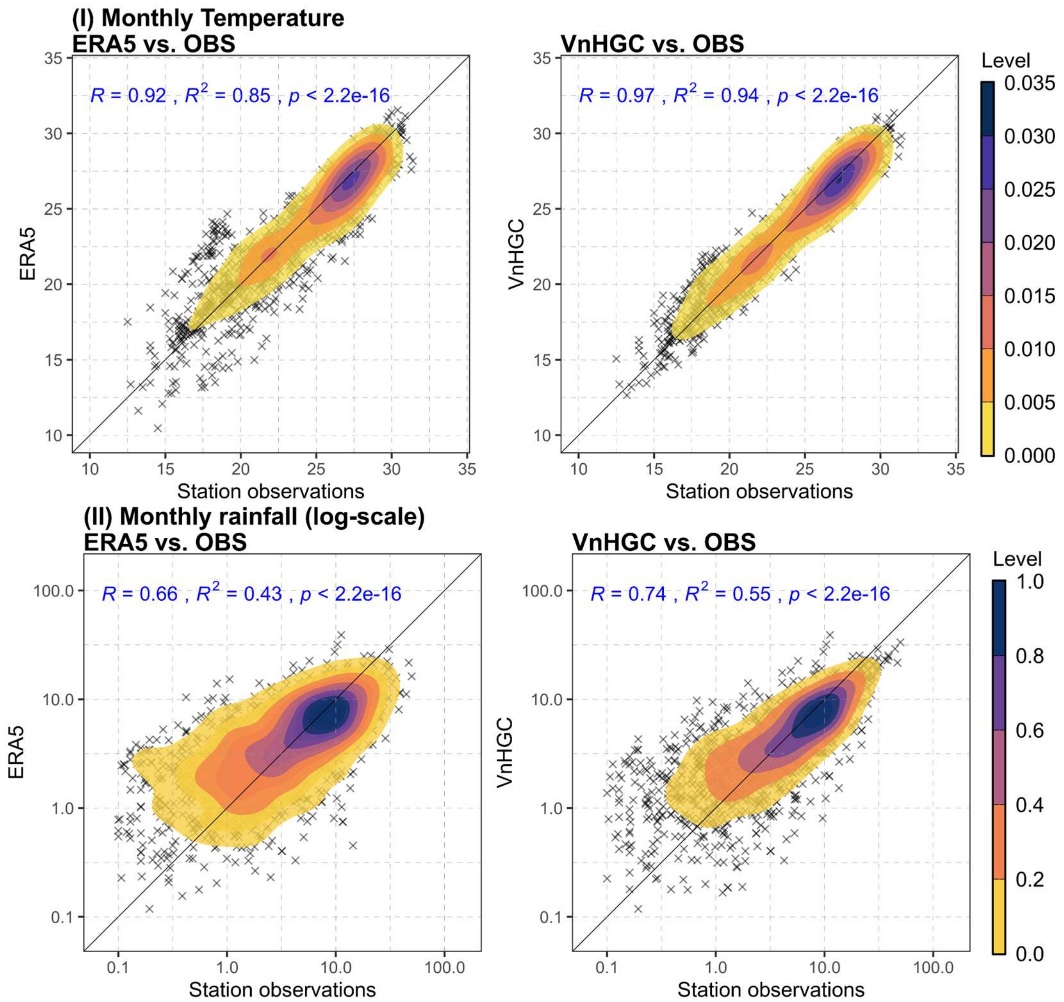


Source: Authors' own calculation, original.

Next, we interpolated ERA5 and VnHGC data to the locations of each station for the 1940–1953 period using the nearest-neighbor method to further investigate the impact of OI. Similar to Figure 6, Figure 9 presents pairwise comparisons between station data and each gridded dataset for all months of the study period. However, unlike Figure 6, where each panel contains over one million data points, each panel in Figure 9 includes only approximately 5,000 points, representing roughly 30 stations across 12 months over 14 years. This smaller sample size, coupled with the sparse station network of the period, already suggests that the evaluation may be less robust compared to the later period.

The dataset for this period was constructed at a monthly scale, which has smoothed out much of the noise inherent in daily variability and extreme events. Consequently, while Figure 9 shows improvements in VnHGC compared to ERA5 for both temperature and precipitation, the signals of improvement are less pronounced and weaker than those observed for the 1961–1980 period (Figure 6). This contrast highlights the combined impact of temporal resolution and observational data availability on the effectiveness of OI.

Figure 9. Scatter-plots comparing monthly means of (I) temperature (°C) and (II) rainfall (mm/day) observations with nearest-neighbor interpolated data from ERA5 (left panels) and VnHGC (right panels) during the period 1949-1953.



Note: The shaded area indicates the density of points. For rainfall results, x- and y-axis are log-transformed for better visualization. Source: Authors' own calculation, original.

3. Summary and conclusions

In this study, we gathered historical data from various sources to create a gridded dataset for Vietnam. This dataset, named VnHGC, encompasses information on temperature and precipitation for two distinct periods: 1940–1953, with monthly data and a horizontal resolution of 25 km; and 1961–1980, featuring daily data and a horizontal resolution of 10 km.

The optimal interpolation (OI) technique, with several adjustments tailored to the characteristics of the station observations we collected, was employed to develop these datasets. The analysis results unequivocally demonstrate the efficacy and appropriateness of OI for our study. The background datasets — ERA5-LAND (used for the daily dataset spanning 1961–1980) and ERA5 (used for the monthly dataset covering 1940–1953) — already exhibited relatively high quality and exhibited a close alignment with station observations. However, they were further optimized by incorporating additional station data.

Overall, the study's findings are promising and provide a strong basis to expand and enhance this research. Several future directions are planned. First, we intend to continue digitizing additional data, not only for the missing period of 1954–1960 — for which we will have data published in the *Résumé mensuel du temps* published from 1949 to 1973 (Thomas and Nguyen, 2024c), but also by incorporating more stations with usable records within the study domain. This effort will be reported in a separate study. Second, while the OI method has proven effective and justified our initial hypotheses, it was initially chosen for its computational efficiency. In the future, we will consider more advanced methods that explicitly account for temporal variability to further improve the dataset. Lastly, since the pre-1961 dataset is currently constructed only at a monthly scale, we plan to explore statistical models, including potentially machine learning techniques, to downscale the monthly data to a daily resolution, making use of the historical daily station data already collected (Thomas et al., 2025), thereby ensuring consistency across the dataset.

In addition to these future plans, one limitation of this study is the absence of comparison with other existing datasets. Satellite products like Tropical Rainfall Measuring Mission (TRMM; Liu et al., 2012) and Climate Prediction Center Morphing technique (CMORPH; Joyce et al., 2004), which have demonstrated reliability across various regions worldwide, could serve as valuable benchmarks for further enhancing VnHGC. Unfortunately, our study period spans the distant past, when satellites and many other gridded datasets were not yet available. As mentioned previously, it is worth noting that a parallel effort to develop a more recent gridded dataset (1980–present) is currently underway in collaboration with IMHEN, using the same OI methodology but based on a much denser station network. The comparisons between the latter dataset and other high-quality satellite-based and reanalysis products will facilitate a more rigorous validation and refinement of both the algorithm and the dataset.

Finally, the development of VnHGC creates new opportunities for future research. With reliable, observation-based gridded data for Vietnam's pre-satellite period, researchers will have a robust reference basis for evaluating regional climate models and reanalysis products. The dataset will also make it possible to explore long-term climate variability and change, overcoming earlier limitations in quantification. In addition, VnHGC can support more applied studies, such as reconstructing past droughts and floods, examining compound weather and climate hazards, and assessing historical impacts on sectors such as agriculture and water resources. By filling a long-standing observational gap, VnHGC strengthens the foundation for advancing a wide range of climate-related scientific and practical applications in Vietnam.

Bibliography

Aybar, C., Fernández, C., Huerta, A., Lavado, W., Vega, F., & Felipe-Obando, O. (2019). Construction of a high-resolution gridded rainfall dataset for Peru from 1981 to the present day. *Hydrological Sciences Journal*, 65(5), 770–785.

Bhattacharyya, S., Sreekesh, S., & King, A. (2021). Characteristics of extreme rainfall in different gridded datasets over India during 1983–2015. *Atmospheric Research*, 267, 105930.

Bui-Manh, H., Peng, C., Fu, Y., Dinh, D., Lin, N., & Yen, M. (2021). The second rainy stage onset in the central highlands of Vietnam. *Geophysical Research Letters*, 48(10).

Bui-Minh, T., Doan, Q-V,, Nguyen, K-C., Phan-Van, T., Trinh-Tuan, L., Cong, T., & Trinh-Minh, N. (2024) Determining the onset of summer rainfall over Vietnam using self-organizing maps. *Climate Dynamics*, 62(9), 9189–9206.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2(1), 1-21.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., . . ., & Thépaut, J. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049.

Hiep, N. H., Luong, N. D., Nga, T. T. V., Hieu, B. T., Ha, U. T. T., Du Duong, B., Long, V. D., Hossain, F., & Lee, H. (2018). Hydrological model using ground- and satellite-based data for river flow simulation towards supporting water resource management in the Red River Basin, Vietnam. *Journal of Environmental Management*, 217, 346–355.

Hong, Y., Adler, R. F., Hossain, F., Curtis, S., & Huffman, G. J. (2007). A first approach to global runoff simulation using satellite rainfall estimation. *Water Resources Research*, 43(8).

Huang, B., Liu, C., Banzon, V. F., Freeman, E., Graham, G., Hankins, W., Smith, T. M., & Zhang, H.-M.. (2020). NOAA 0.25-degree Daily Optimum Interpolation Sea Surface Temperature (OISST), Version 2.1 [Dataset]. *NOAA National Centers for Environmental Information*. https://doi.org/10.25921/RE9P-PT57. Accessed [October 22, 2025].

Joyce, R. J., Janowiak, J. E., Arkin, P. A., & Xie, P. (2004). CMORPH: A Method that Produces Global Precipitation Estimates from Passive Microwave and Infrared Data at High Spatial and Temporal Resolution. *Journal of Hydrometeorology*, 5, 487–503.

Kalnay, E. (2003). Atmospheric Modeling, Data Assimilation and Predictability. *Cambridge University Press, Cambridge*.

Kubota, T., Aonashi, K., Ushio, T., Shige, S., Takayabu, Y. N., Kachi, M., Arai, Y., Tashima, T., Masaki, T., Kawamoto, N., Mega, T., Yamamoto, M. K., Hamada, A., Yamaji, M., Liu, G., & Oki, R. (2020). Global Satellite Mapping of Precipitation (GSMAP) products in the GPM era. In *Satellite precipitation measurement: Volume 1* (pp. 355-373). Cham: Springer International Publishing.

Le, X., Kim, Y., Van Binh, D., Jung, S., Nguyen, D. H., & Lee, G. (2024). Improving rainfall-runoff modeling in the Mekong river basin using bias-corrected satellite precipitation products by convolutional neural networks. *Journal of Hydrology*, 630, 130762.

Liu, Z., Ostrenga, D., Teng, W., & Kempler, S. (2012). Tropical Rainfall Measuring Mission (TRMM) Precipitation Data and Services for Research and Applications. *Bulletin of the American Meteorological Society*, 93, 1317–1325.

Lussana, C., Saloranta, T., Skaugen, T., Magnusson, J., Tveito, O. E., & Andersen, J. (2018). seNorge2 daily precipitation, an observational gridded dataset over Norway from 1957 to the present day. *Earth System Science Data*, 10(1), 235–249.

Main, L., Sparrow, S., Antje Weisheimer, & Wright, M. (2024). Skilful probabilistic medium-range precipitation and temperature forecasts over Vietnam for the development of a future dengue early warning system. *Meteorological Applications*, 31(4).

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J. (2021). ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9), 4349–4383.

Ngo-Duc T., Matsumoto, J., Kamimera, H., & Bui, H.-H. (2013) Monthly adjustment of Global Satellite Mapping of Precipitation (GSMaP) data over the Vu Gia–

Thu Bon River Basin in Central Vietnam using an artificial neural network. *Hydrological Research Letters*, 7(4), 85-90.

Nguyen, N. Y., Anh, T. N., Nguyen, H. D., & Dang, D. K. (2024). Quantile mapping technique for enhancing satellite-derived precipitation data in hydrological modelling: a case study of the Lam River Basin, Vietnam. *Journal of Hydroinformatics*, 26(8), 2026–2044.

Nguyen-Xuan, T., Ngo-Duc, T., Kamimera, H., Trinh-Tuan, L., Matsumoto, J., Inoue, & Phan-Van, T. (2016) The Vietnam Gridded Precipitation (VnGP) Dataset: Construction and Validation. SOLA, 12,291–296.

Nguyen-Xuan, T., Nguyen-Le, D., & Ngo-Duc, T. (2025). Performance evaluation of RegCM5 in long-term convection-permitting simulations for high-urbanized areas in Vietnam. *Vietnam Journal of Earth Sciences*, *47*(3), 411-429.

Nodzu, M.I., Matsumoto, J., Trinh-Tuan, L., & Ngo-Duc, T. (2019) Precipitation estimation performance by Global Satellite Mapping and its dependence on wind over northern Vietnam. *Progress in Earth and Planetary Science*, *6*(1), 58.

Parkes, B., Higginbottom, T. P., Hufkens, K., Ceballos, F., Kramer, B., & Foster, T. (2019). Weather dataset choice introduces uncertainty to estimates of crop yield responses to climate variability and change. *Environmental Research Letters*, 14(12), 124089.

Roversi, G., Pancaldi, M., Cossich, W., Corradini, D., Nguyen, T. T. N., Nguyen, T. V., & Porcu, F. (2024). The extreme rainfall events of the 2020 typhoon season in Vietnam as seen by seven different precipitation products. *Remote Sensing*, 16(5), 805.

Serrano-Notivoli, R., Begueria, S., Saz, M. Á., Longares, L. A., & De Luis, M. (2017). SPREAD: a high-resolution daily gridded precipitation dataset for Spain – an extreme events frequency and intensity overview. *Earth System Science Data*, 9(2), 721–738.

Thomas, F., & Nguyen, T. N.-M. (2024a). Climate data rescue 5: Annuaire statistique de l'Indochine (ASI) [Dataset]. DataSuds. https://doi.org/10.23708/Y82VD1

Thomas, F., & Nguyen, T. N.-M. (2024b). Climate data rescue 3: Bulletin Mensuel des Observations (BMO) [Dataset]. DataSuds. https://doi.org/10.23708/WH1UPA

Thomas, F., Nguyen, T.N.-M. (2024c). Climate data rescue 7: Résumé mensuel du temps (RMT), [Dataset]. DataSuds. https://doi.org/10.23708/M9BSO7

Thomas, F., Ngo-Duc, T., & Nguyen, N.M.T. (2025). The climate archives of Indochina and their contribution in understanding climate change in Vietnam and Southeast Asia. *AFD research papers*, accepted.

Tran, V. N., Nguyen, H. D., Van Khuong, H., Dao, H. B., Le, Q. H. M., Nguyen, C. Q., & Nguyen, G. T. (2025). Reconstructing Long-Term Daily Streamflow data at the discontinuous monitoring station in the ungauged transboundary basin using machine learning. *Water Resources Management*, 39, 3327–3348.

Tran-Anh, Q., Ngo-Duc, T., Espagne, E., Trinh-Tuan, L. (2023). A 10-km CMIP6 downscaled dataset of temperature and precipitation for historical and future Vietnam. *Scientific Data*, 10:257. https://doi.org/10.1038/s41597-023-02159-2

Trinh, T., Do, N., Nguyen, V. T., & Carr, K. (2021). Modeling high-resolution precipitation by coupling a regional climate model with a machine learning model: An application to Sai Gon–Dong Nai Rivers Basin in Vietnam. *Climate Dynamics*, 57(9), 2713–2735.

Trinh-Tuan, L., Matsumoto, J., Tangang, F.T., Juneng, L., Cruz, F., Narisma, G., Santisirisomboon, J., Phan-Van, T., Gunawan, D., Aldrian, E., & Ngo-Duc, T. (2019a). Application of Quantile Mapping Bias Correction for Mid-Future Precipitation Projections over Vietnam. SOLA, 15,1–6. https://doi.org/10.2151/sola.2019-001

Trinh-Tuan, L., Matsumoto, J., Ngo-Duc, T., Nodzu, M. I., & Inoue, T. (2019b). Evaluation of satellite precipitation products over Central Vietnam. *Progress in Earth and Planetary Science*, 6(1), 54. https://doi.org/10.1186/s40645-019-0297-7

Tuan, B. M. (2019). Extratropical forcing of submonthly variations of rainfall in Vietnam. *Journal of Climate*, 32(8), 2329–2348.

Van, T. N., Dang, M. N., Quang, V. D., Minh, T. B., Van, K. M., Van, K. D., Thanh, T. T., Hoang, D. T., Thi, T. T., Huu, Q. N., Xuan, T. L., & Duy, H. T. (2023). Orographic effect and the opposite trend of rainfall in central Vietnam. *Advances in Meteorology*, 1–12.

Vu, M. T., Raghavan, S. V., & Liong, S. Y. (2012). SWAT use of gridded observations for simulating runoff – a Vietnam river basin study. *Hydrology and Earth System Sciences*, 16(8), 2801–2811.

List of acronyms and abbrevations

AFD	Agence française de développement
ASI	Annuaire Statistique de l'Indochine
C3S	Copernicus Climate Change Service
CCDS	Copernicus Climate Data Store
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	fifth generation of ECMWF reanalysis
IMHEN	Institute of Meteorology, Hydrology and Climate Change
OI	Optimal Interpolation
VnGC	Vietnam Gridded Climate Dataset
VnGP	Vietnam Gridded Precipitation
VnHGC	Vietnam Historical Gridded Climate
VNMHA	Vietnam Meteorological Hydrological Administration

Agence française
de développement
5, rue Roland Barthes
75012 Paris l France
www.afd.fr

What is AFD?
Éditions Agence française de développement publishes analysis and research on sustainable development issues. Conducted with numerous partners in the Global North and South, these publications contribute to a better understanding of the challenges faced by our planet and to the implementation of concerted actions within the framework of the Sustainable Development Goals.
With a catalogue of more than 1,000 titles and an average of 80 new publications published every year, Éditions Agence française de développement promotes the dissemination of knowledge and expertise, both in AFD's own publications and through key partnerships. Discover all our publications in open access at editions.afd.fr.
Towards a world in common.

Publication Director Rémy Rioux
Editor-in-Chief Thomas Melonio

Legal deposit 4th quarter 2025
ISSN 2492 - 2846

Rights and permissions
Creative Commons license
Attribution - No commercialization - No modification
https://creativecommons.org/licenses/by-nc-nd/4.0/

BY NC ND

Graphic design MeMo, Juliegilles, D. Cazeils
Layout PUB
Printed by the AFD reprography service

To browse our publications:
https://www.afd.fr/en/ressources-accueil