REVIEW
SPECIAL TOPIC: Paleoclimate modelling

SCIENCE CHINA Earth Sciences
earth.scichina.com link.springer.com
https://doi.org/10.1007/s11430-025-1810-2

Progress and prospects of paleoclimate data assimilation

Liang NING[1], Jian LIU[1,2*], Zhengyu LIU[3], Fangmiao XING[1], Fen WU[1], Mi YAN[1], Zilu MENG[4], Kefan CHEN[1], Yanmin QIN[1], Weiyi SUN[1] & Qin WEN[1]

[1] State Key Laboratory of Climate System Prediction and Risk Management; Key Laboratory for Virtual Geographic Environment, Ministry of Education; State Key Laboratory Cultivation Base of Geographical Environment Evolution and Regional Response of Jiangsu Province; Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application; School of Geography, Nanjing Normal University, Nanjing 210023, China
[2] Jiangsu Provincial Key Laboratory for Numerical Simulation of Large Scale Complex Systems, School of Mathematical Science, Nanjing Normal University, Nanjing 210023, China
[3] Department of Geography, The Ohio State University, Columbus 43210, USA
[4] Department of Atmospheric and Climate Science, Seattle 98195, USA

* Corresponding author (email: jliu@njnu.edu.cn)

Received 21 August 2025; Revised 8 December 2025; Accepted 22 December 2025; Published online 9 February 2026

Abstract Reconstructing paleoclimate characteristics and understanding their evolution rules is a key question in Earth system science and global change research. It helps clarify the historical position of the modern warming period, understand the features and mechanisms of climate change under warming backgrounds, and thereby improve the accuracy of future climate projections. Proxy records and numerical modeling are two primary approaches in current paleoclimate study. As an emerging methodology, paleoclimate data assimilation effectively integrates paleoclimate proxy records with numerical simulations, combining their advantages to enhance the accuracy of paleoclimate reconstructions. This paper systematically reviews recent progresses in paleoclimate data assimilation. It first outlines the historical developments of major paleoclimate data assimilation methods, discusses their advantages, disadvantages, and applicability, and highlights recent improvements such as the application of machine learning methods and the developments of online assimilation. Then, it introduces applications of paleoclimate data assimilation according to different typical paleoclimate periods, particularly addressing challenges associated with various types of proxy data, and summarizes currently available open-source datasets and algorithm platforms. A specific case study is presented to illustrate the application of assimilating oxygen isotope simulations. Finally, the paper discusses unresolved issues and challenges in paleoclimate data assimilation studies, and outlines potential directions for future research.

Keywords Paleoclimate data assimilation, Proxy records, Model simulations, Proxy system model, Uncertainty

1. Introduction

Paleoclimate research contributes to understanding climate change mechanisms, identifying key climatic factors influencing the environment, and evaluating contemporary climate numerical models. This, in turn, helps improve predictions of future climate and environmental changes and supports sustainable developments (Wang, 2022). Paleoclimate proxy records and model simulations are two primary research methods in paleoclimate studies. Each of them has its own advantages: paleoclimate proxy records represent the historical evidence of paleoclimate, while model simulations incorporate dynamic mechanisms.

Paleoclimate proxy records refer to various biological, physical, and chemical indicators used to reconstruct paleoclimates. For the last two millennia, proxy datasets include the PAGES2k dataset (Past Global Changes 2k) (PAGES2k Consortium, 2017) and the Mann09 dataset (Mann et al., 2009). Holocene proxy datasets include Temperature 12ka (Kaufman et al., 2020), the Arctic Holocene Proxy Climate Database (Sundqvist et al., 2014), and the LegacyClimate 1.0 dataset (Herzschuh et al., 2023), etc. In addition, there are various integrated datasets dedicated to specific proxy types, such as the CoralHydro2k dataset for coral records (Walter et al., 2023) and the SASIL dataset for stalagmite records (Comas-Bru et al., 2020). These datasets have significantly contributed to the paleoclimate reconstruction across different typical periods.

For paleoclimate modeling, the Paleoclimate Modelling Intercomparison Project (PMIP) has long been dedicated to understanding past climate states and their responses to external forcings through model simulations. Currently, the latest PMIP4 results have been released, which include transient simulations of the past millennium (past1000), as well as equilibrium simulations for the Mid-Holocene (MH), the Last Glacial Maximum (LGM), the Last Interglacial (LIG), the mid-Pliocene, and the Early Eocene (EECO), the Paleocene-Eocene Thermal Maximum (PETM), and the pre-PETM (Kageyama et al., 2018). Continuous transient simulations include multi-member ensemble simulations for the past millennium, such as the Community Earth System Model-Last Millennium Ensemble (CESM-LME) (Otto-Bliesner et al., 2016), and the Nanjing Normal University Last Two Millennium Simulation (NNU-2ka) (Wang et al., 2016). For the Holocene, continuous transient simulations include the Nanjing Normal University Holocene Simulation (NNU-Holocene) (Wan et al., 2020), the HT-11.5ka experiment by the Institute of Atmospheric Physics, Chinese Academy of

Citation: Ning L, Liu J, Liu Z, Xing F, Wu F, Yan M, Meng Z, Chen K, Qin Y, Sun W, Wen Q. 2026. Progress and prospects of paleoclimate data assimilation. Science China Earth Sciences, https://doi.org/10.1007/s11430-025-1810-2

Sciences (Tian et al., 2020), the experiment based on MPI-ESM by the Max Planck Institute for Meteorology in Germany (Bader et al., 2020), the experiment based on HadCM3 by the Hadley Centre in the UK (Hopcroft and Valdes, 2021, 2022), and the experiment based on EC-Earth by Stockholm University in Sweden (Zhang et al., 2021). Continuous transient simulations since the LGM include the Simulation of the Transient Climate of the Last 21,000 Years (TraCE-21ka) (Liu et al., 2009) and the latest Isotope-enabled Simulation of the Transient Climate of the Last 21,000 Years (iTraCE-21ka) (He et al., 2021). Additionally, there are accelerated transient simulations covering the past 300,000 years (Xie et al., 2019; Yan et al., 2023). These simulations contribute to understanding the mechanisms of paleoclimate changes across different typical periods.

However, each of these two approaches has its own drawbacks. For instance, paleoclimate proxy records are often point-based, with spatiotemporal discontinuities and uneven distributions. They are subject to dating errors, and the climatic indications of some proxies are ambiguous or controversial (Chen et al., 2023). Moreover, proxies primarily reflect surface variables such as temperature or precipitation. Additionally, paleoclimate proxy records involve two types of errors, i.e., instrumental errors arising from human factors or uncertainties in measurement and analytical instruments during sampling and analysis, and spatial representativeness errors due to mismatches between the spatial scale represented by the proxy records and that of climate model grids (Li, 2013).

On the other hand, model simulations often reflect the internal variability of the models themselves, which may not accurately capture the phases of true internal variability in paleoclimates. They also exhibit uncertainties in the magnitudes of responses to external forcings. Furthermore, although paleoclimate modeling aims to better represent external forcings and feedbacks, current state-of-the-art models are primarily developed and calibrated for future climate predictions and projections rather than being specifically designed or optimized for paleoclimate studies (Kageyama et al., 2018), which may introduce biases in simulation accuracy.

Therefore, integrating the advantages of both approaches and compensating for their respective drawbacks are essential in paleoclimate research to derive more accurate paleoclimate characteristics and rules. This is the original motivation of paleoclimate data assimilation (von Storch et al., 2000). Paleoclimate data assimilation provides a mathematical framework to extract useful information from proxy records and simulations. Proxy records offer evidence of climate changes, while simulations provide a physically constrained framework based on dynamical equations. By quantitatively estimating errors in both proxy records and simulations, this approach constrains paleoclimate model runs (or directly adjusts simulation results), yielding more accurate and spatiotemporally continuous paleoclimate reconstructions (Hakim et al., 2013).

Thus, paleoclimate data assimilation shares a common objective with modern climate data assimilation, i.e., utilizing spatially discontinuous observations to generate spatially regularized reanalysis data, especially for variables not directly observed. However, paleoclimate data assimilation faces unique challenges, primarily because it deals with proxy records whose physical meanings are often unclear. This introduces complexities in assimilation steps, such as the design of observational operators and error quantification. For example, stalagmite $\delta^{18}O$ is a proxy for monsoon intensity, but its specific climatic interpretation (e.g., circulation strength, moisture source variations) remains debated. To reconstruct monsoon precipitation using such proxy, it is necessary to develop nonlinear proxy system models (PSMs) with the aid of isotope-enabled simulations, posing challenges distinct from modern data assimilation.

Due to its unique advantages, many controversies between proxy records and simulations, such as the Holocene temperature conundrum (Liu et al., 2014), can also be addressed through paleoclimate data assimilation. By providing continuous spatial fields, assimilation enables more precise depictions of the spatial distribution of global and regional climate changes, thereby enhancing the understanding on responses of both global and regional climate to different forcings. Furthermore, assimilation can contribute to key paleoclimate scientific questions, such as climate sensitivity, by delivering more accurate results that offer more reliable references for future projections. Consequently, it facilitates a deeper understanding of the historical position and impacts of the modern warming period, promoting the integration of paleo- and modern climate and environment research (Wang, 2022).

Previous studies have provided detailed reviews of the principles, methods, and applications of paleoclimate data assimilation (e.g., Fang and Li, 2016; Zhang et al., 2025; Tierney et al., 2025b). In recent years, significant advancements have been made in assimilation algorithms and "online" assimilation techniques (Sun et al., 2022; Meng and Hakim, 2024). Notably, the inclusion of oxygen isotope simulations has improved nonlinear proxy system models (PSMs) and online assimilation, attracting widespread attention in the paleoclimate community. Therefore, this paper will briefly review the historical developments of the principles, methods, and applications of paleoclimate data assimilation, with a focus on recent innovations and techniques (e.g., online assimilation strategies). It will then discuss the theoretical, technical, and data-related challenges currently faced by paleoclimate data assimilation studies. Finally, the paper will also explore the potential applications of paleoclimate data assimilation to key scientific questions and outline future research directions for priority investigation.

2. Paleoclimate data assimilation methods and evolutions

In simple terms, the fundamental concept of paleoclimate data assimilation is to use proxy records to constrain model simulations, combining the results from previous time steps to produce an optimal estimate of the current climate state. Its core principle is based on traditional Bayesian theory:

$$P(\mathbf{x} \mid \mathbf{y}) \propto P(\mathbf{y} \mid \mathbf{x}) \cdot P(\mathbf{x}) \tag{1}$$

where $\mathbf{x}$ represents the reconstructed climate variables, $\mathbf{y}$ represents the proxy records, $P(\mathbf{x})$ stands for the prior probability provided by the model simulations, $P(\mathbf{y} \mid \mathbf{x})$ is the likelihood function, indicating the probability of proxy data given the climate state, and $P(\mathbf{x} \mid \mathbf{y})$ refers to the posterior probability of the climate variables obtained after assimilation, i.e., the assimilated results. As shown in the formula, the final assimilated results depend on both the prior distribution and the likelihood. Efforts in paleoclimate data assimilation thus focus on estimating these

two components and solving for the posterior distributions. Therefore, paleoclimate data assimilation primarily consists of four components, i.e., paleoclimate model simulations, proxy records, PSMs, and assimilation algorithms.

The specific process (as illustrated in Fig. 1) is as follows: for a given time step and specific variables in paleoclimate data assimilation, paleoclimate simulations are first used to generate the required data, serving as the prior estimate. Then, the prior estimate is transformed into the proxy data space through PSMs. The difference between the actual proxy record value and the proxy record value estimated via the PSMs at that time step is calculated, referred as the "innovation" in paleoclimate data assimilation. Finally, the weight of the innovation is calculated based on the model covariance and proxy record error, and is applied to update the prior estimate, thereby obtaining the posterior estimate (Talagrand, 1997). This process will be repeated for next time step, ultimately producing assimilated results that incorporate both historical climate evidence and physical mechanisms.

In practical paleoclimate data assimilation studies, the primary distinctions lie in the estimation and optimization of prior information. Based on their chronological development, the mainstream paleoclimate data assimilation algorithms currently include nudging, particle filter, offline ensemble Kalman filter, and recent online assimilation methods (Fig. 2). Brief introductions to the principles and applications of these methods are as follows.

2.1 Nudging

The nudging method is a data assimilation technique that adds a forcing term into the forecast model, and gradually drives the model state towards the observations (Hoke and Anthes, 1976). The formula is as follows:

$$\psi^n = f(\psi^{n-1}) + \alpha H^T(d^{n-1} - H(\psi^{n-1})) + \xi^n \quad (2)$$

where ψ^n represents the model state at time t_n, which is a function f of the state ψ^{n-1} at time t_{n-1}; α is the nudging parameter; H is the operator that transforms the model state into the observation space; d^n is the observational data at time t_n; and ξ^n is random noise.

The primary advantage of the nudging method is its simplicity, straightforwardness, and ease of implementation, while also offering strong constraint effects. However, its drawbacks are equally evident, since nudging can only assimilate variables directly output by the model, requiring the transformation of observational data into model-output variables during assimilation. Moreover, as the formula indicates, the parameter α determines the strength of the nudging effect. Too strong nudging may induce erroneous dynamics due to excessively rapid convergence, while too weak nudging may fail to effectively constrain the results with observations (Dubinkina and Goosse, 2013). Additionally, the selection of α is typically based on empirical approaches, lacking a solid physical foundation.

In terms of application, the nudging method was the earliest

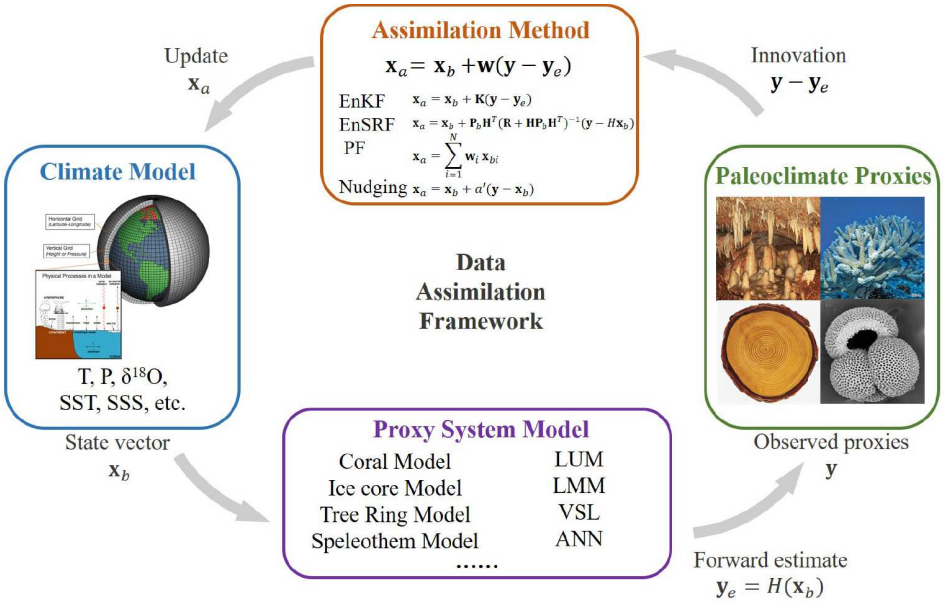

Fig. 1 Conceptual framework for paleoclimate data assimilation.

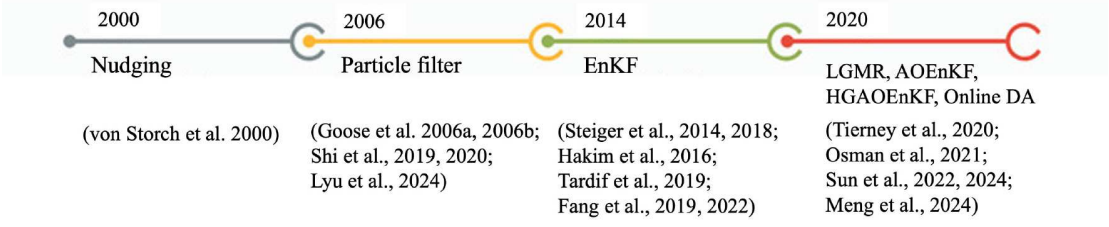

Fig. 2 Evolution of paleoclimate data assimilation methods.

approach used in paleoclimate data assimilation. von Storch et al. (2000) employed nudging to reconstruct the North Atlantic Oscillation (NAO) index during historical periods, and found that the assimilated results better reproduced the true variations of historical climate compared to model simulations. When comparing three assimilation methods, including nudging, for reconstructing Northern European climate over the past millennium, Widmann et al. (2010) concluded that nudging could effectively reconstruct climate changes over the past millennium but struggled to capture target modes of variability that differed from the model's internal variability. Dubinkina and Goosse (2013) compared the performance of three methods, i.e., nudging, particle filter, and particle filter integrated with nudging, in reconstructing high-latitude Southern Hemisphere climate over the past 150 years. They found that pure nudging underperformed the other two methods in assimilating variables with no direct observations (such as sea surface salinity, SSS), as it failed to reflect oceanic dynamical processes. Since then, there are relatively limited applications of nudging in paleoclimate data assimilation research.

2.2 Particle filter

The basic idea of a particle filter is to approximate the posterior probability distribution of the state by weighting a set of random model sample particles based on Bayesian likelihood estimation. The formula for calculating the posterior probability distribution (Dubinkina and Goosse, 2013) is as follows:

$$p(\psi^n|d^n) = \sum_{i=1}^{M} \omega_i^n \delta\left(\psi^n - \psi_i^n\right) \tag{3}$$

where δ is the kernel density, and ω_i^n is the weight of each particle, calculated using the following formula:

$$\omega_i^n = K^{-1} p\left(d^n \mid \psi_i^n\right) \tag{4}$$

where K is the normalization coefficient, and $p\left(d^n \mid \psi_i^n\right)$ represents the likelihood of the observations given the model state.

The advantage of a particle filter is that it does not require the prior distribution to be Gaussian (Dubinkina and Goosse, 2013), nor does it assume a linear relationship between observations and prior estimates. Its drawbacks include high requirements on the quantity and quality of observational data, as well as a tendency for weights to concentrate on a small number of particles.

In the early applications of particle filter, simplified versions were often employed, where only the simulation closest to the observations was selected as the optimal particle to serve as the initial condition for the next assimilation step (Goosse et al., 2006; Widmann et al., 2010). For instance, Goosse et al. (2006) applied a simplified particle filter method to simulate Northern Hemisphere climate over the past millennium. Using only a small number of particles (30 particles) and simple weight calculations, they were able to generate climate states consistent with the records. In subsequent studies, the particle filter was compared with nudging, such as in the two aforementioned stduies (Widmann et al., 2010; Dubinkina and Goosse, 2013). Regarding the performance of particle filter alone, Widmann et al. (2010) applied a simplified particle filter to assimilate temperature data in Northern Europe, successfully reproducing multi-decadal temperature variability despite using only 11 particles (simulation results). This suggests that the common issue of particle degeneracy did not arise. Similarly, Dubinkina and Goosse (2013) found that particle filter effectively reconstructed variables without direct observations, such as SSS, particularly when combined with nudging. In terms of improving particle filter, Dubinkina et al. (2011) and Annan and Hargreaves (2012) incorporated residual resampling methods into the algorithm. Their results demonstrated that the standard particle filter with residual resampling significantly outperformed simplified particle filter in terms of assimilation accuracy.

Besides the algorithm improvements, beyond the reconstruction of temperature and circulation fields, particle filter has been applied in recent years to precipitation reconstruction in broader regions such as East Africa, East Asia, and South America (Klein and Goosse, 2018; Shi et al., 2019; Lyu et al., 2024), achieving favorable results. Notably, a recent study (Lyu et al., 2024) applied particle filter to reconstruct South American monsoon precipitation and circulation fields, using over 600 particles. It was found that particle filter effectively captures the nonlinear dynamic relationship between $\delta^{18}O$ and precipitation, leading to better assimilated results.

2.3 Offline ensemble Kalman filter

The offline ensemble Kalman filter (EnKF) has been widely used in the field of paleoclimate data assimilation in recent years (Hakim et al., 2016; Tardif et al., 2019; Tierney et al., 2020; Li et al., 2024; Wu et al., 2025). The core idea of EnKF is to update the expected value at each time step using paleoclimate proxy records, while assigning weights by comparing the error of the proxy records with the observational error covariance. The ensemble concept is reflected in the estimation of the background error covariance matrix based on statistical characteristics. The specific formula is as follows:

$$\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{b}} + \mathbf{K}[\mathbf{y} - H(\mathbf{x}^{\mathrm{b}})] \tag{5}$$

where $\mathbf{x}^{\mathrm{a}}$ is the assimilated result; $\mathbf{x}^{\mathrm{b}}$ is the prior estimate, typically sampled from a static source or obtained through conditional sampling, such as existing climate model simulations; $\mathbf{y}$ is the proxy records; H is the PSM that transforms the prior estimate into the proxy space; $\mathbf{y} - H(\mathbf{x}^{\mathrm{b}})$ characterizes the difference between the observed data and the prior estimate; and $\mathbf{K}$ is the Kalman gain matrix. $\mathbf{K}$ is used to assign weight to $\mathbf{y} - H(\mathbf{x}^{\mathrm{b}})$ and transform it into the state ($\mathbf{x}^{\mathrm{b}}$) space. The calculation formula is as follows:

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathrm{T}}[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1} \tag{6}$$

where $\mathbf{B}$ is the covariance matrix of the prior estimate, $\mathbf{R}$ is the error covariance matrix of the observations, and $\mathbf{H}$ is a linear PSM.

The main advantages of the offline EnKF method are the high accuracy under the given assumptions, a relatively straightforward solution process, ease of parallel computation, and convenient system implementation. These features have contributed to its widespread application in the field of paleoclimate data assimilation in recent years. Its primary drawbacks include the assumptions that the errors in prior estimates and observations follow Gaussian distributions, and that the relationships

between observations and model results are linear.

In terms of applications, early studies conducted a series of idealized experiments on the applications of the EnKF in paleoclimate data assimilation. For example, Huntley and Hakim (2010) tested the sensitivity of the EnKF to the distribution of observation sites and found that when the number of sites is limited, assimilated results based on a small number of well-distributed sites are comparable to the results based on a large number of randomly distributed sites. Pendergrass et al. (2012) demonstrated that assimilation skill significantly improves compared to statistically based reconstructions under two conditions, i.e., when the model's forecast skill exceeds the temporal resolution of the proxy records, and when climate covariance is strongly correlated with the mean state. Steiger et al. (2014) applied the EnKF to temperature reconstruction over the past millennium, compared it with traditional principal component analysis (PCA) methods, and found that the EnKF results are more reliable in terms of spatial features, particularly in regions with sparse proxy data. Subsequently, Hakim et al. (2016) and Tardif et al. (2019) further applied the EnKF method to produce the Last Millennium Reanalysis (LMR). In recent years, newly developed reanalysis datasets for the last two millennia (Hu et al., 2024; Wu et al., 2025), Holocene temperature reconstructions (Erb et al., 2022), and Last Glacial Maximum Reanalysis (LGMR, Tierney et al., 2020; Osman et al., 2021) have all been based on the EnKF method.

2.4 Online data assimilation

Since offline assimilation methods construct prior distributions based solely on a static source, such as existing climate simulations, they lack memory of previous climate states. Therefore, many researchers have begun exploring online assimilation methods to address this limitation (Perkins and Hakim, 2017, 2020). For example, Perkins and Hakim (2017, 2020) used a linear inverse model (LIM) in the assimilation of annual temperature and circulation fields to perform online predictions of the posterior, generating prior distributions for the next assimilation step. Their findings indicate that online assimilation methods outperform offline ones, with improvements largely attributed to the dynamical constraints of the coupled ocean-atmosphere system. Compared to earlier online assimilation methods that often relied on LIM, Meng and Hakim (2024) developed an online EnKF assimilation system based on a deep learning model, and further reconstructed monthly tropical Pacific sea surface temperature (SST), meridional and zonal wind stress, and upper-ocean temperatures across seven layers. They found that because deep learning models can capture more nonlinear relationships between current and future climate states and retain greater prediction capacity, they yield more accurate forecasts than traditional LIM. Moreover, these improvements vary by region and variable, primarily manifesting in extratropical zonal wind stress and SST, equatorial ocean temperatures, and the thermocline in the central Pacific. In terms of precipitation assimilation, due to the lower memory precipitation, the skill of online precipitation assimilation remains below that of temperature and circulation fields (Perkins and Hakim, 2020), similar to offline assimilation.

Due to the longer memory of the ocean, online assimilation can transfer oceanic memory to the atmosphere, thereby improving atmospheric assimilation, especially for the past millennium, where proxy records are predominantly derived from terrestrial indicators (Perkins and Hakim, 2020; Meng and Hakim, 2024; Meng et al., 2025). However, for longer time scales, several questions remain to be explored, e.g., to what extent can oceanic memory enhance the prediction skill of online assimilation, how reliable is it for predictions of decadal and longer scale variability, what role can deep-sea proxies like foraminifera play in improving long-scale prediction skill, and to what degree does prediction skill depend on the model used? Furthermore, current online assimilation efforts have largely focused on the last two millennia, as longer time periods require substantially greater computational resources, making simple climate models more feasible for such applications. Nevertheless, with advances in computational power in the future, Earth system models may also become viable for online assimilation over longer time periods.

3. Applications of paleoclimate data assimilations

In recent years, the applications of paleoclimate data assimilation in reconstructing climates in different typical periods have yielded numerous significant results. Previous studies have provided detailed summaries of these applications (Zhang et al., 2025). Here, a brief review is offered from the perspective of technical details.

3.1 Last two millennia

The last two millennia are the most mature period for the application of paleoclimate data assimilation. Since the introduction and subsequent methodological advancements, paleoclimate data assimilation has largely focused on climate reconstruction for this period (e.g., von Storch et al., 2000; Goosse et al., 2010). In recent years, significant progresses have been made in data assimilation for this period, driven by improvements in both proxy records and model simulations (Zhu et al., 2023). Commonly used proxy datasets include the PAGES2k and Mann09 datasets, as well as newer datasets such as CoralHydro2k. Widely utilized simulations include the PMIP past1000 simulations and the CESM-LME simulations. In terms of assimilation methods, early techniques such as nudging, particle filter, and EnKF have all been applied, with particle filter and EnKF being more prevalent in recent years. Furthermore, substantial methodological advancements have been achieved, such as the development of the Analogue Offline Ensemble Kalman Filter (AOEnKF) and Hybrid Gain Analogue Offline Ensemble Kalman Filter (HGAOEnKF) by Sun et al. (2022, 2024). These methods enhance assimilation skills by refining the sampling of prior distributions. Additionally, recent online assimilation approaches (Meng and Hakim, 2024; Meng et al., 2025; Sun et al., 2025) have further contributed to these advancements.

Currently, the major assimilated datasets include the Last Millennium Reanalysis (LMR) (Hakim et al., 2016; Tardif et al., 2019), the Paleo Hydrodynamics Data Assimilation Product (PHYDA) (Steiger et al., 2018), and the Nanjing Normal University Last Two Millennia Reanalysis (NNU-2ka Reanalysis) (Hu et al., 2024; Wu et al., 2025). In addition to conventional variables such as temperature, precipitation, and circulation fields, these datasets also include indices like the Palmer Drought Severity Index (PDSI), the Intertropical Convergence Zone (ITCZ), El Niño-Southern Oscillation (ENSO), the Pacific Decadal Oscilla-

tion (PDO), and the Atlantic Multidecadal Oscillation (AMO).

Because of the maturity of assimilation techniques for the last two millennia, paleoclimate data assimilation has been applied not only for reconstructing climate characteristics but also for analyzing the mechanisms of multi-scale climate variability (Zhu et al., 2022). For instance, Erb et al. (2020) reconstructed drought and circulation fields in the United States over the past millennium and revealed that internal variability, rather than external forcings, dominated multi-year droughts. Lyu et al. (2024) reconstructed South American monsoon intensity over the past millennium and found a centennial-scale strengthening during the transition from the Medieval Climate Anomaly (MCA) to the Little Ice Age (LIA), which was linked to a southward shift of the Atlantic Intertropical Convergence Zone (ITCZ) and an intensification of the Pacific Walker Circulation. Additionally, Fang et al. (2022) assimilated proxy records from tree rings, ice cores, lake sediments, and historical documents in the Arctic to reconstruct the Arctic amplification index over the past millennium, and indicated that the AMO dominated its multi-decadal variations, while anthropogenic greenhouse gases drove its centennial-scale weakening since the Industrial Revolution.

Overall, assimilation of temperature and circulation fields over the last two millennia has reached a relatively mature stage. However, challenges remain in precipitation assimilation, as precipitation exhibits greater spatial heterogeneity and local variability compared to temperature (Hancock et al., 2023). Furthermore, the relationships between precipitation and proxy records, as well as the underlying mechanisms, are more complex (Wu et al., 2025). Additionally, due to the abundance of proxy records and more developed PSMs, the last two millennia also serve as a testing ground for future advancements of assimilation methods.

3.2 Holocene

Compared to the last two millennia, the applications of paleoclimate data assimilation in the Holocene and earlier periods is relatively limited, with the EnKF method being the primary assimilation algorithm used. For the Holocene period, Erb et al. (2022) employed the EnKF method to reconstruct spatially and temporally continuous temperature variations. The proxy records used were from the Temperature 12k dataset (Kaufman et al., 2020), which includes indicators from lake sediments, marine sediments, peat, ice cores, and stalagmites. The simulations were derived from the LGM transient experiment using the HadCM3 model (Snoll et al., 2022) and the TraCE-21ka experiment based on the CCSM3 model (Liu et al., 2014). The findings indicate that the Mid-Holocene temperature was the highest in the pre-industrial era, approximately 0.09 °C higher than that of the past millennium. This result is lower than previous Holocene reconstructions (Marcott et al., 2013; Kaufman et al., 2020) but higher than other assimilated results (Osman et al., 2021). Additionally, Erb et al. (2022) also investigated the influence of seasonality on Holocene temperature trends and found that even when accounting for summer biases across all records, the discrepancies between proxy records and simulations could not be fully explained.

3.3 Since the LGM

Tierney et al. (2020) and Osman et al. (2021) applied the EnKF method to reconstruct temperature changes since the LGM. The proxy records they used consisted of marine geochemical indicators representing SST, including $\delta^{18}O$, Mg/Ca, $U_{37}^{K'}$, TEX_{86}. The simulations were four snapshot experiments conducted using the iCESM. The assimilated results show that the global mean temperature during the LGM decreased by −6.1 °C (with a 95% confidence interval of −6.5 to −5.7 °C), corresponding to an estimated climate sensitivity of 3.4 °C (with a 95% confidence interval of 2.4–4.5 °C) (Tierney et al., 2020). The primary drivers of temperature changes since the LGM were radiative forcings induced by ice sheets and greenhouse gases, followed by variations in the Atlantic Meridional Overturning Circulation (AMOC) and seasonal solar radiation (Osman et al., 2021). Annan et al. (2022) also employed the EnKF method, utilizing multi-model results from PMIP and three sets of gridded SST and surface air temperature datasets to reconstruct LGM sea surface and surface air temperatures. Their results indicate that the global mean temperature anomaly during the LGM relative to the pre-industrial era was −4.5±0.9 °C. The discrepancy between this result and that of Tierney et al. (2020) primarily stems from differences in prior selection. Consequently, they recommend using multi-model ensembles as reliable prior estimates, provided that the range of simulations comprehensively and realistically captures the main sources of uncertainty.

3.4 Deep-time data assimilation

In recent years, paleoclimate data assimilation has also been applied to the reconstruction of deep-time climates. For instance, focusing on the Paleocene-Eocene Thermal Maximum (PETM, 56 Ma), Tierney et al. (2022) utilized the EnKF method to reconstruct the climate state during the PETM. In addition to the four marine geochemical indicators used in the LGMR assimilation, terrestrial proxy indicators such as $MBT_{Me}^{5'}$ were incorporated. The simulations were derived from a set of Early Eocene experiments conducted using the iCESM model. The assimilated results indicate that the global mean temperature anomaly during the PETM was 5.6 °C (with a 95% confidence interval of 5.4–5.9 °C), corresponding to an estimated climate sensitivity of 6.5 °C (with a 95% confidence interval of 5.7–7.4 °C) (Tierney et al., 2022).

Li et al. (2024) also employed the EnKF method to reconstruct carbon cycle perturbations during the PETM. The proxy indicators used included deep-sea sedimentary $CaCO_3$ and SST proxies ($\delta^{18}O$, Mg/Ca, and TEX_{86}). The simulations were a 100-member ensemble of experiments conducted with the cGENIE model. The assimilated results show that atmospheric CO_2 increased from 890 ppm(1 ppm=1 μL/L) to 1980 ppm, seawater pH decreased by 0.46, and the saturation state of calcium carbonate in seawater declined from 10.2 to 3.8.

The Pliocene (5.33–2.58 Ma) is the most recent geological period when atmospheric CO_2 concentrations approached 400 ppm. Tierney et al. (2025a) used the EnKF method to reconstruct the climate state of the Pliocene, utilizing SST proxies ($\delta^{18}O$, Mg/Ca, and TEX_{86}) as indicators. The simulations include 14 PlioMIP2 experiments, 2 Pliocene sensitivity experiments based on CESM2, and 21 Pliocene-like experiments based on CESM1. The assimilated results indicate that the mid-Pliocene warming was approximately 4.1 °C (with a 95% confidence interval of 3.0–5.3 °C), corresponding to an estimated climate

sensitivity of 4.8 °C (with a 95% confidence interval of 2.6–9.9 °C). Additionally, the equatorial Pacific SST gradient exhibited an El Niño-like pattern, with lower salinity over the North Pacific but higher salinity over the North Atlantic.

Judd et al. (2024) further employed the EnKF method to reconstruct global mean temperatures during the Phanerozoic (485 Ma) (PhanDA). The proxy records consisted of marine geochemical indicators representing SST, including $\delta^{18}O$, Mg/Ca, $U_{37}^{K'}$, TEX_{86}. The simulations were 80-member snapshot experiments conducted using the iCESM. The assimilated results show that the global mean temperature varied between 11 to 36 °C, corresponding to an estimated climate sensitivity of approximately 8 °C. Moreover, atmospheric CO_2 concentration was identified as the primary driver of global mean temperature variations throughout the Phanerozoic.

Overall, assimilation methods for the last two millennia are relatively mature and diverse, including particle filter, EnKF and its variants, as well as various recent online assimilation techniques. In contrast, for the Holocene and earlier periods, assimilation primarily relies on the EnKF method. In terms of initial conditions, assimilation for the last two millennia and the Holocene mainly utilizes results from transient simulations, while for the LGM and earlier periods, snapshot simulation results are predominantly used. Regarding proxy records, assimilation for the last two millennia focuses on tree rings, coral $\delta^{18}O$, and Sr/Ca ratios, etc. For the Holocene, assimilation relies more on stalagmite $\delta^{18}O$, lake and marine sediments, whereas for the LGM and earlier periods, marine geochemical indicators are the primary proxies used.

3.5 Existing assimilation datasets and platforms

For the last two millennia, in terms of assimilation datasets, Hakim et al. (2016) combined the PAGES 2ka reconstruction dataset with simulations of the last millennium to produce the LMR, which includes temperature, precipitation, and circulation fields. Steiger et al. (2018) further reconstructed the Paleo Hydrodynamics Data Assimilation Product (PHYDA), encompassing drought-wetness indices and circulation fields for the last two millennia. Hu et al. (2024) developed the NNU-2ka Reanalysis. Erb et al. (2022) reconstructed Holocene temperature data, while the Tierney team produced temperature change datasets for the LGM (LGMR), the Paleocene–Eocene Thermal Maximum (PETM), the Pliocene, and the Phanerozoic. All these datasets have been made openly accessible.

Among these datasets, the assimilated results for the last two millennia include not only conventional variables, such as temperature and precipitation, but also circulation fields at different altitudes. In contrast, the assimilated results for the Holocene and earlier periods primarily focus on temperature reconstructions. Therefore, reconstructing circulation fields for these earlier periods represents one of the future research directions. However, achieving this goal requires a deeper understanding of the mechanisms driving climate change during these periods, as well as improvements in model simulations.

Regarding assimilation platforms, many of the current assimilation algorithms are open-source, primarily based on the EnKF or its variants in Python or MATLAB. Key platforms include the LMR toolkit (Hakim et al., 2016), which corresponds to the LRM, its modified version LMR Turbo (LMRt) (Zhu et al., 2021), and the more recent Climate Field Reconstruction (cfr) toolkit (Zhu et al., 2024), all of which are developed in Python. The research team led by Jessica Tierney at the University of Arizona has also developed the DASH software package (King et al., 2023), which corresponds to the Last Glacial Maximum Reanalysis (LGMR) and is based on MATLAB. The AOEnKF and HGAOEnKF software packages developed by Lili Lei's team at Nanjing University (Sun et al., 2022, 2024) are also based on MATLAB. Additionally, the deepDA package developed by Mingsong Li's team at Peking University (Li et al., 2024) is implemented in Python. The core algorithms of all these software packages are based on the EnKF or its improved variants.

3.6 Case study

Oxygen isotope $\delta^{18}O$, as a proxy indicator preserved in multiple archives, is widely used in paleoclimate reconstruction across various time periods. Previous reconstructions based on oxygen isotopes predominantly relied on linear regression models, often neglecting the physical mechanisms underlying the relationship between $\delta^{18}O$ and climate variables (Liu et al., 2023). Today, with the availability of isotope-enabled simulations, the physical mechanisms driving $\delta^{18}O$ variations can be better understood, enabling assimilation of oxygen isotopes from various archives using nonlinear PSMs (Tierney et al., 2020; Lyu et al., 2024). Lyu et al. (2024) demonstrated that incorporating $\delta^{18}O$ into the assimilation process improves the reconstruction of South American monsoon variability compared to earlier datasets such as LMR and PHYDA. So, how does assimilation skill differ between nonlinear PSMs and traditional linear PSMs? Below, we specifically compare the differences in SST assimilation using linear and nonlinear PSMs for coral $\delta^{18}O$.

Fig. 3A and 3B show the Niño3.4 indices from assimilated results based on linear and nonlinear PSMs, respectively, compared with observations. The results indicate that both methods yield assimilated results with high correlations with observations (r=0.82, 0.81, p<0.01), with no significant difference between them. Fig. 3C and 3D illustrate the covariance spatial fields used in the two assimilation methods, using the Palmyra Island record as an example. The spatial correlation fields between SST and $\delta^{18}O$ at Palmyra Island and global SST resemble the typical ENSO pattern, indicating that both variables effectively represent SST in the Niño3.4 region and can be used for assimilating the Niño3.4 index. The assimilated results show minimal differences between the two methods, with the nonlinear PSM assimilation results exhibiting a wider distribution. The coefficient of efficiency for the nonlinear PSM is 0.35, which is lower than the 0.55 for the linear PSM. This suggests that the two methods perform similarly in reproducing Niño3.4 index variability, though the nonlinear PSM results cover a broader range.

This indicates that the assimilation performance of nonlinear PSMs based on isotope-enabled simulations is comparable to that of traditional linear PSMs. Consequently, applying nonlinear PSMs constructed from simulations to the assimilation of proxy indicators such as stalagmite oxygen isotopes holds promising prospects for the future. Additionally, these approaches can be used to refine linear PSMs for proxy indicators, such as foraminiferal isotopes and tree-ring isotopes, as well as to validate assimilated results.

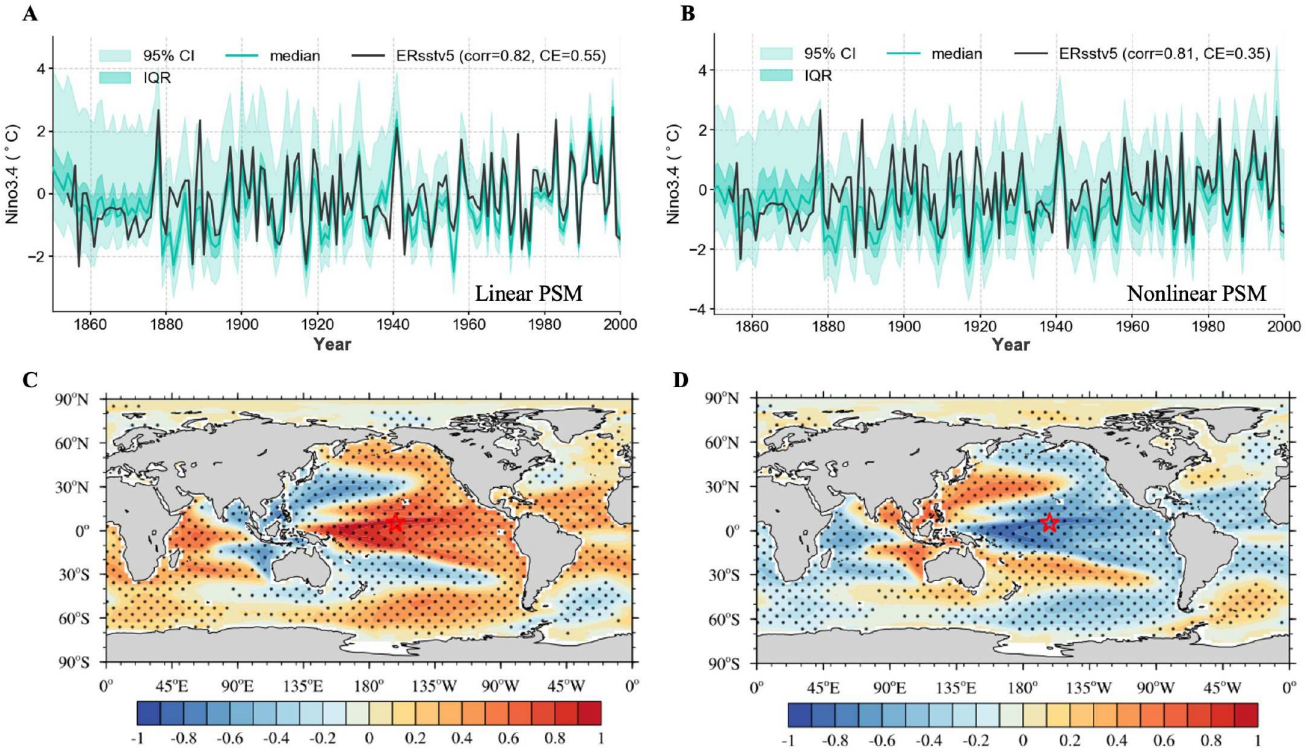

Fig. 3 Comparison of the reconstructed Niño3.4 indices based on the linear PSM (A) and the nonlinear PSM (B) during the observational period, along with the spatial correlation fields representing the covariance matrices used in assimilation (C, D). In panel D, seawater $\delta^{18}O_{sw}$ was converted to coral $\delta^{18}O_c$ before calculating the correlation coefficients.

4. Prospects

Since the concept of paleoclimate data assimilation was introduced in 2000, it has made significant progresses in paleoclimate research. This paper systematically reviews the development of paleoclimate data assimilation, its applications in different typical periods, and the key scientific questions it addresses. It can be said that paleoclimate data assimilation has greatly enhanced our understanding of the spatiotemporal characteristics and evolutions of climate in various historical periods. In periods with well-established applications of paleoclimate data assimilations, such as the last two millennia, assimilation methods can also be used to improve our understanding of the mechanisms behind climate change. Currently, research plans related to paleoclimate data assimilation within several international scientific programs are also being actively prepared, such as the Paleoclimate Data Assimilation Model Intercomparison Project (Paleo-DA MIP) in CMIP7 and the Paleoclimate Reanalysis and Integration of Synthesis and Models (PRISM) project in PAGES2k.

Indeed, paleoclimate data assimilation research is still in its early stages. Even the currently widely used offline EnKF method, while simplified and effective, still has several aspects that urgently require resolution or further refinement. The following sections will outline prospects for key scientific questions that can be addressed by paleoclimate data assimilation, its advantages and limitations, potential improvements, and future research directions.

4.1 Key scientific questions that can be addressed by paleoclimate data assimilation

How to better reconstruct the characteristics and rules of paleoclimate by integrating proxy records and model simulations is the primary scientific question that paleoclimate data assimilation needs to address. Beyond the fundamental reconstructions of multiple scale climate changes, paleoclimate data assimilation is also applied to resolve controversies between proxy records and simulations, as well as among proxy records themselves. A typical example is the Holocene temperature conundrum. Erb et al. (2022) argue that the Mid-Holocene temperature was higher than the past millennium, whereas the assimilated results of Osman et al. (2021) and Bova et al. (2021) do not show a pronounced Mid-Holocene warming period. Erb et al. (2022) also conducted sensitivity experiments, revealing that while seasonal biases in proxy indicators may reflect summer temperature trends, the potential impact of such biases is insufficient to align the reconstructed global mean temperature with the warming trend observed in transient simulations. In contrast, Osman et al. (2021) suggest that the overemphasis on sparse proxy data from the Southern Hemisphere may partially induce the Mid-Holocene warming period, but this only accounts for a fraction of the warm anomaly. Further research is needed to reconcile the discrepancies between proxy records and simulations, particularly regarding whether paleoclimate data assimilation can contribute to this issue.

Recently, Hao et al. (2025) reconstructed Holocene temperatures using global marine sediment proxies and found significant spatial heterogeneity in Mid-Holocene temperature anomalies. Specifically, winter and annual mean temperatures were higher in Europe and high-latitude Eurasia, while other regions exhibited lower temperatures. The discrepancies were attributed to a cold bias in high-latitude regions due to biases in vegetation and sea ice feedbacks in the models, as well as a warm bias due to

the concentration of proxy records in Europe.

Additionally, recent alkenone-based reconstructions in mid-latitude Eurasia (Jiang et al., 2024) also highlight spatial heterogeneity in Holocene temperature trends, showing cooling in northeastern China but warming in southwestern Siberia. This provides crucial insights about the regional localization of covariance matrices when assimilating proxy records at different regions. Furthermore, a recent study (Liu et al., 2025) indicates that proxy records may underestimate temperature seasonality, causing reconstructed Holocene temperature trends to be dominated by summer temperatures. Whether data assimilation can reduce this uncertainty by integrating simulations to produce a more reliable Holocene temperature trend remains an important question for further exploration.

In addition to reconstructing climate characterisitics, data assimilation is being applied to mechanism studies, particularly for the well-established period of the last two millennia. Currently, assimilation for the last two millennia has successfully reconstructed circulation fields, while assimilation for other typical periods remains focused on reconstructing the spatio-temporal characteristics of variables such as temperature. For example, studies of the past millennium have examined multi-scale monsoon precipitation or drought variations and their driving mechanisms (Erb et al., 2020; Lyu et al., 2024). This is because the PSMs linking proxy indicators to circulation changes during the last two millennia are relatively well-understood. Key factors include the availability of proxy records for calibration of PSMs during the instrumental period, and the mature climatic interpretation of proxy indicators for this period.

In terms of algorithms, circulation fields reconstructed using particle filter exhibit greater consistency with temperature and precipitation variations, whereas other methods, which assimilate circulation fields and temperature/precipitation separately, show less consistency. However, the reliability of particle filter for reconstructing circulation fields depends heavily on the model's ability to accurately simulate the underlying mechanisms.

Beyond temperature, precipitation, and circulation fields, paleoclimate data assimilation can also reconstruct a broader range of climatic and environmental variables. For instance, Li et al. (2024) reconstructed carbon cycling and carbonate saturation states during the PETM, providing valuable insights into ocean acidification and seawater carbonate saturation. This approach also offers a novel perspective for reconstructing the climatic and environmental conditions of even older geological periods.

Additionally, quantitatively differentiating the contributions from internal variability and external forcings, another key scientific question in paleoclimate research, is also an important aspect of mechanism studies facilitated by paleoclimate data assimilation. Previous research often relied on single-forcing sensitivity experiments to differentiate these contributions, but such methods suffer from model dependency. In contrast, paleoclimate data assimilation, by integrating results from multiple models, can provide a more robust estimate of the responses to external forcings, thereby contributing to addressing this issue.

For example, under specific external forcings, early assimilation methods yielded results that did not exceed the range of internal variability from the model simulations (Widmann et al., 2010). In the assimilation process, sources of internal variability include randomly selected initial conditions from multi-model ensembles and higher-frequency variability beyond the signal (or noise) in proxy records. Sources of the response to external forcings encompass the simulated responses from multi-model ensembles and the signal within the proxy records. By quantitatively assessing the reliability of these two sources, a more accurate representation of the responses to external forcings and internal variability can be achieved, thereby enabling a better quantitative differentiation of their relative contributions to climate change.

In terms of contributing to future climate projections, climate sensitivity is also one of the key scientific questions frequently addressed in paleoclimate data assimilation. Climate sensitivity is defined as the change in the Earth's surface temperature when atmospheric CO_2 concentration doubles. Its calculation method (Tierney et al., 2020) is as follows:

$$\mathrm{ECS} = \frac{\Delta \mathrm{GMST}}{\Delta R} \times F_{2\times \mathrm{CO}_2} \tag{7}$$

where ΔGMST is the change in global mean surface temperature for the typical period obtained through assimilation, ΔR is the radiative forcing for that period, and $F_{2\times \mathrm{CO}_2}$ is the radiative forcing resulting from a doubling of CO_2 concentration. In the calculation process, ΔR and $F_{2\times \mathrm{CO}_2}$ are estimated from model simulations, while the assimilated results provide the value of ΔGMST. Based on the LGMR, Tierney et al. (2020) estimated the climate sensitivity to be 3.4 °C (with a 95% confidence interval of 2.4–4.5 °C). Tierney et al. (2025a) estimated the climate sensitivity based on Pliocene temperature assimilation to be 4.8 °C (with a 90 % confidence interval of 2.6–9.9 °C). Judd et al. (2024) estimated the climate sensitivity based on Phanerozoic temperature assimilation to be approximately 8 °C. These results show that the value of climate sensitivity increases with the temperature of the typical periods, and the variation in confidence intervals indicates that the distribution of assimilated results expands as the period extends further back in time. Furthermore, the uncertainty in estimating radiative forcing for the typical period also significantly impacts the calculation of climate sensitivity, underscoring the need for further improvement in reconstructing external forcings and enhancing the accuracy of model feedbacks for these periods.

4.2 Limitations of paleoclimate data assimilations and potential improvements

In recent years, researchers have made significant improvements to address the limitations of various components in paleoclimate data assimilation, such as PSMs and assimilation algorithms. These improvements are reflected in better selection of priors for assimilation, more accurate construction of relationships between observed and simulated variables, and advances in online assimilation techniques.

As previously discussed, the quantification of uncertainties is one of the primary challenges in paleoclimate data assimilation, with significant difficulties in uncertainty estimations of both proxy records and model simulations. When calculating the uncertainty of proxy records, it is essential to account for statistical uncertainty inherent in the records, dating errors, instrumental measurement errors, spatial representativeness errors, and representation errors of the PSM. Quantifying these uncertainties requires in-depth investigation and close collaboration with experts in reconstruction fields. For estimating the

uncertainty of model simulations, a large number of independent samples of long-term climate means are needed, specifically, long-term simulations that fall within the observational error range but vary in external forcings, boundary conditions, and model parameters. Currently, computational resources are insufficient to meet this demand. Moreover, most current assimilation studies rely on results from a single model. Future assimilation efforts should incorporate multi-model simulations to better quantify the uncertainties associated with model outputs.

Regarding the quality control and error estimation of proxy records, although more proxy records can provide more paleoclimate information and contribute to more accurate reconstructions, simply adding new proxy records into the assimilation process is not sufficient. Instead, appropriate quality control must be applied, followed by the quantification of their uncertainties. Then, an objective assessment of the weight that they account for in the assimilation process should be made.

During the assimilation process, it is essential to first quantify the contributions of proxy records from different regions for assimilations targeting different periods, scales, and variables (Wu et al., 2025). Regarding dating errors, the high accuracy of tree-ring dating can be leveraged in the last two millennia to partially calibrate proxy data with annual or higher resolutions, such as coral records (Hu et al., 2024). For longer time scales, in the absence of proxy records with precise dating for calibration, integrating dating errors into the assimilation process requires thorough collaboration with experts in reconstruction fields to incorporate their empirical knowledge. Additionally, beyond the commonly used proxy records, other less frequently applied records, particularly qualitative materials such as historical documents, also need to be considered for integration into assimilation. To achieve this, the error estimation methods for different types of proxy records should account for characteristics like resolution and dating accuracy, enabling the effective fusion of diverse data sources.

In terms of PSM construction and refinement, the models for tree-rings, coral oxygen isotopes, marine foraminifera, and $U_{37}^{K'}$ are now relatively well-developed. However, PSMs for stalagmite oxygen isotopes, pollen, and other proxies require further development and improvement (Ning et al., 2025a). In particular, PSMs linking these proxies to precipitation remain in their preliminary stages. In this context, with the aid of oxygen isotope-enabled simulations, PSMs for isotope-related proxies such as stalagmite oxygen isotopes can be effectively constructed (Ning et al., 2025b). Nonetheless, the differences between these nonlinear operators and traditional linear operators need to be systematically compared to evaluate their respective advantages and limitations. Additionally, machine learning methods have already been applied to construct nonlinear PSMs, and hold potential for further refinement in the future (Fang and Li, 2019; Wei et al., 2024). For example, Fang and Li (2019) developed a nonlinear PSM for tree-ring width using an artificial neural network approach, demonstrating that its assimilation performance surpasses the linear regression and the VS-Lite model. This also confirms the feasibility of applying machine learning methods to the assimilation of other variables in the future.

In terms of validating assimilated results, if the assimilated period overlaps with the instrumental era, observational data can be used for validation, with common metrics including the correlation coefficient, root mean square error, and coefficient of efficiency (Zhang et al., 2025). However, for longer time periods lacking observational data, a common approach is to withhold 25% of the proxy records for independent validation (Hakim et al., 2016; Tierney et al., 2020; Osman et al., 2021; Wu et al., 2025). This method, however, incorporates the uncertainties of the proxy records themselves, which may affect the objectivity of the validation. Recent studies have attempted to use independent observational data, such as borehole temperatures, for validation (Meng et al., 2025). Therefore, exploring more objective methods for validating assimilated results remains an important direction for future research.

4.3 Future research directions of paleoclimate data assimilations

Building on the aforementioned improvements, paleoclimate data assimilation can be further developed in several areas, including online paleoclimate assimilation, paleoclimatic dynamical constraints, and the applications of deep learning and big data.

4.3.1 Paleoclimate online data assimilation

In terms of future developments in assimilation algorithms, besides improvements to traditional offline assimilation methods (Sun et al., 2022, 2024), online assimilation has recently emerged (Perkins and Hakim, 2020; Meng and Hakim, 2024; Sun et al., 2025). It is important to note, however, that paleoclimate online data assimilation differs from online integration assimilation used in modern climate studies. The main distinction lies in the fact that the initial field at the current time step is generated from the state vectors of the preceding 12 months using machine learning methods (Meng and Hakim, 2024). Compared to conventional offline assimilation methods and online assimilation methods based on LIM, machine learning-based online assimilation has demonstrated higher accuracy for the last two millennia, particularly under conditions where proxy records are sparse (Sun et al., 2025).

Current online assimilation efforts largely focus on monthly scale reconstructions for the last two millennia. Key challenges include whether temporal resolution can be further increased, whether the methods can be extended to longer time periods, and how to enhance computational efficiency while leveraging the advantages of online assimilation in constructing initial fields. Moreover, deep learning-based online assimilation algorithms still suffer from signal attenuation issues, requiring methods such as inflation to increase initial ensemble perturbations to enhance ensemble spread (Meng and Hakim, 2024; Sun et al., 2025).

4.3.2 Applications of paleoclimatic dynamical constraints in paleoclimate data assimilation

As the primary dynamical constraint in the assimilation process, the refinement and application of paleoclimate dynamics play a crucial role in advancing assimilation methods. First, further clarifying the climatic interpretations of proxy indicators is essential. Currently, the climatic interpretations of many proxies are often inferred based on correlation coefficients, while the underlying physical processes remain poorly understood. For example, while ice core $\delta^{18}O$ from the South American Andes is commonly assumed to reflect ENSO variability, this correlation actually arises from the influence of orbital scale SST anomalies

in the eastern equatorial Pacific on mid-tropospheric water vapor $\delta^{18}O$ (Liu et al., 2023).

Additionally, variations in stalagmite $\delta^{18}O$ in East Asia, which may reflect local precipitation, moisture source shifts, and upstream depletion effects, can be quantitatively disentangled using isotope-enabled simulations (Ning et al., 2025b), thereby improving the accuracy of PSMs. By leveraging models to determine the climatic interpretation of proxies, we can quantitatively distinguish the representations of a single proxy to different climate variables. This capability not only expands the application of assimilation to local variables but also enables the assimilation of large-scale SST or circulation fields through teleconnection relationships . Furthermore, when integrating different types of proxy records in assimilation, the quantitative representation of climate variables by various proxies should also be taken into account.

However, it is important to note that when applying dynamical constraints, the covariance matrix of simulations should not be used indiscriminately. Models themselves contain errors; for instance, Sanchez et al. (2021) have found that common model biases, such as the double ITCZ bias, significantly impact ENSO reconstruction from coral records in the SPCZ region, necessitating error correction before assimilation. Similarly, biases in simulating teleconnections also affect assimilated results . On the other hand, the strength of teleconnection influences from large-scale circulation fields varies across different characteristic periods (Ning et al., 2025a). Therefore, dynamical constraints in assimilation must also evolve over time. Currently, while transient simulations are used for periods such as the last two millennia (Hakim et al., 2016; Wu et al., 2025) and the Holocene (Erb et al., 2022), longer-term assimilations (e.g., Tierney et al., 2020; Li et al., 2024) predominantly rely on snapshot simulations. In future research, the application of transient simulations for longer periods (such as iTraCE experiments) could improve the selection of initial fields in assimilation by better capturing the effects of external forcings on the climate system. Thus, while the usage of transient simulations has the potential to improve long-term assimilations, further comparative validation is required.

Furthermore, the dynamical constraints derived from paleoclimate data assimilation results can assist in optimizing the selection of sites for proxy record collection. Huntley and Hakim (2010) conducted a series of sensitivity experiments and found that when observations are sparse, the location of sites is more critical to the accuracy of assimilated results than the number of sites. This suggests that site selection for proxy record collection can be optimized by referencing assimilated results. Wu et al. (2025) also observed that, with a comparable number of proxy records, the spatial distribution of these records significantly impacts assimilated results. Recent improvements in assimilation algorithms have also focused on enhancing performance under conditions of sparse proxy data (Sun et al., 2025). Therefore, sensitivity experiments can be designed to quantitatively evaluate the contribution of site information from proxy data to climate reconstructions. In summary, compared to direct model simulations, paleoclimate data assimilation incorporates the climatic interpretation of proxy records, enabling it to provide more informed theoretical guidance for selecting proxy collection sites based on the type and attributes of the proxies (Fang and Li, 2016). However, most of these findings are based on assimilation in the last two millennia. Similar sensitivity analyses for longer time periods remain to be conducted.

4.3.3 Applications of deep learning and big data in paleoclimate data assimilation

Machine learning and big data have been widely applied in Earth sciences, offering new approaches to many complex problems. In paleoclimate data assimilation research, deep learning methods were first utilized in constructing PSMs. Their advantage lies in not requiring a clear understanding of the physical mechanisms linking proxy indicators to climate variables (Fang and Li, 2019). However, their construction demands large amounts of data for training, which currently limits their application primarily to proxy indicators such as tree rings (Fang and Li, 2019) and corals (Wei et al., 2024), as these records have sufficient data during the observational period. For other proxy indicators with only limited temporal coverage overlapping with the observational record, constructing PSMs based on deep learning remains challenging. Nonetheless, this lack of explicit physical understanding may introduce potential biases into assimilated results. Moreover, similar to linear PSMs, deep learning-based PSMs also carry the risk of overfitting (Fang et al., 2022). Additionally, the impact of proxy record quality on model construction requires thorough evaluation.

Besides deep learning, other machine learning methods also hold significant potential for application in paleoclimate data assimilation. For example, causal inference methods (Su et al., 2023) can be used to clarify the genuine causal relationships between proxy indicators and climate variables, moving beyond the commonly used linear correlations. This could enhance the reliability and interpretability of reconstruction processes. Transfer learning offers considerable advantages in constructing PSMs, as it can overcome limitations such as scarce training data and high computational costs. However, it still carries the risk of overfitting. Before assimilating proxy records with ambiguous climatic interpretations, causal discovery algorithms can be employed to determine whether a specific climate variable directly drives changes in the proxy indicator or whether multiple variables involve confounding factors. This approach would help in selecting the most reliable proxy indicators for assimilation. Additionally, physics-informed neural networks can leverage known physical laws to guide the training of neural networks. This enables the construction of PSMs that adhere to physical laws, even in regions where proxy records are sparse.

The construction of data-driven climate models is also one of the primary applications of machine learning in the field of data assimilation. For instance, Meng and Hakim (2024) introduced deep learning into paleoclimate data assimilation to develop an online assimilation system. They found that this approach yields more accurate reconstructions of tropical upper-ocean temperatures compared to traditional LIM. Sun et al. (2025) further advanced this by building an online assimilation framework for the last two millennia based on deep learning-based networks and an integrated hybrid EnKF. Their results demonstrate that, through techniques such as inflated observational errors, this method achieves higher accuracy than conventional LIM-based online assimilation and offline assimilation, particularly in earlier periods when proxy records are relatively sparse.

In addition, machine learning has also been applied to model error estimation. For example, Peng et al. (2024) developed a model error estimation method based on convolutional neural networks to quantify errors arising from inaccurate model

parameters and initial conditions. When applied to data assimilation with simplified models, this method demonstrated effective correction of model errors, suggesting its potential for future application in paleoclimate data assimilation.

Acknowledgement

This work was supported by the National Key Research and Development Program of China (Grant No. 2023YFF0804704), and the National Natural Science Foundation of China (Grant Nos. 42130604, 42575050, 42475051 & 42575051).

Compliance and ethics

The authors declare no conflict of interest.

References

Annan J D, Hargreaves J C. 2012. Identification of climatic state with limited proxy data. Clim Past Discuss, 8: 481–503

Annan J D, Hargreaves J C, Mauritsen T. 2022. A new global surface temperature reconstruction for the Last Glacial Maximum. Clim Past, 18: 1883–1896

Bader J, Jungclaus J, Krivova N, et al. 2020. Global temperature modes shed light on the Holocene temperature conundrum. Nat Commun, 11: 4726

Bova S, Rosenthal Y, Liu Z, et al. 2021. Seasonal origin of the thermal maxima at the Holocene and the last interglacial. Nature, 589: 548–553

Chen F, Duan Y, Hao S, et al. 2023. Holocene thermal maximum mode versus the continuous warming mode: Problems of data-model comparisons and future research prospects. Sci China Earth Sci, 66: 1683–1701

Comas-Bru L, Rehfeld K, Roesch C, et al. 2020. SISALv2: A comprehensive speleothem isotope database with multiple age-depth models. Earth Syst Sci Data, 12: 2579–2606

Dubinkina S, Goose H, Sallaz-Damaz Y, et al. 2011. Testing a particle filter to reconstruct climate changes over the past centuries. Int J Bifurcat Chaos, 21: 3611–3618

Dubinkina S, Goosse H. 2013. An assessment of particle filtering methods and nudging for climate state reconstructions. Clim Past, 9: 1141–1152

Erb M P, Emile-Geay J, Hakim G J, et al. 2020. Atmospheric dynamics drive most interannual U.S. droughts over the last millennium. Sci Adv, 6: eaay7268

Erb M P, McKay N P, Steiger N, et al. 2022. Reconstructing Holocene temperatures in time and space using paleoclimate data assimilation. Clim Past, 18: 2599–2629

Fang M, Li X. 2016. Paleoclimate data assimilation: Its motivation, progress and prospects. Sci China Earth Sci, 59: 1817–1826

Fang M, Li X. 2019. An artificial neural network-based tree ring width proxy system model for paleoclimate data assimilation. J Adv Model Earth Syst, 11: 892–904

Fang M, Li X, Chen H W, et al. 2022. Arctic amplification modulated by Atlantic Multidecadal Oscillation and greenhouse forcing on multidecadal to century scales. Nat Commun, 13: 1865

Goosse H, Renssen H, Timmermann A, et al. 2006. Using paleoclimate proxy-data to select optimal realisations in an ensemble of simulations of the climate of the past millennium. Clim Dyn, 27: 165–184

Goosse H, Crespin E, de Montety A, et al. 2010. Reconstructing surface temperature changes over the past 600 years using climate model simulations with data assimilation. J Geophys Res, 115: D09108

Hakim G J, Annan J, Brönnimann S, et al. 2013. Overview of data assimilation methods. PAGES news, 21: 72–73

Hakim G J, Emile-Geay J, Steig E J, et al. 2016. The last millennium climate reanalysis project: Framework and first results. J Geophys Res-Atmos, 121: 6745–6764

Hancock C L, McKay N P, Erb M P, et al. 2023. Global synthesis of regional Holocene hydroclimate variability using proxy and model data. Paleoceanog Paleoclimatol, 38: e2022PA004597

Hao S, Zhang X, Duan Y, et al. 2025. Model seasonal and proxy spatial biases revealed by assimilated mid-Holocene seasonal temperatures. Sci Bull, 70: 2014–2022

He C, Liu Z, Otto-Bliesner B L, et al. 2021. Hydroclimate footprint of pan-Asian monsoon water isotope during the last deglaciation. Sci Adv, 7: eabe2611

Herzschuh U, Böhmer T, Li C, et al. 2023. LegacyClimate 1.0: A dataset of pollen-based climate reconstructions from 2594 Northern Hemisphere sites covering the last 30 kyr and beyond. Earth Syst Sci Data, 15: 2235–2258

Hoke J E, Anthes R A. 1976. The initialization of numerical models by a dynamic relaxation technique. Mon Weather Rev, 104: 1551–1556

Hopcroft P O, Valdes P J. 2021. Paleoclimate-conditioning reveals a North Africa land-atmosphere tipping point. Proc Natl Acad Sci USA, 118: e2108783118

Hopcroft P O, Valdes P J. 2022. Green Sahara tipping points in transient climate model simulations of the Holocene. Environ Res Lett, 17: 085001

Hu W, Ning L, Liu Z, et al. 2024. Reconstructing tropical monthly sea surface temperature variability by assimilating coral proxy datasets. NPJ Clim Atmos Sci, 7: 261

Huntley H S, Hakim G J. 2010. Assimilation of time-averaged observations in a quasi-geostrophic atmospheric jet model. Clim Dyn, 35: 995–1009

Jiang J, Meng B, Wang H, et al. 2024. Spatial patterns of Holocene temperature changes over mid-latitude Eurasia. Nat Commun, 15: 1507

Judd E J, Tierney J E, Lunt D J, et al. 2024. A 485-million-year history of Earth's surface temperature. Science, 385: 1316

Kageyama M, Braconnot P, Harrison S P, et al. 2018. The PMIP4 contribution to CMIP6—Part 1: Overview and over-arching analysis plan. Geosci Model Dev, 11: 1033–1057

Kaufman D, McKay N, Routson C, et al. 2020. A global database of Holocene paleotemperature records. Sci Data, 7: 1–34

King J, Tierney J, Osman M, et al. 2023. DASH: A MATLAB toolbox for paleoclimate data assimilation. Geosci Model Dev, 16: 5653–5683

Klein F, Goosse H. 2018. Reconstructing East African rainfall and Indian Ocean sea surface temperatures over the last centuries using data assimilation. Clim Dyn, 50: 3909–3929

Li M. 2024. mingsongli/deepDA: V1.0.0-Nature Geoscience Release (v1.0.0). Zenodo, doi: 10.5281/zenodo.13777776

Li M, Kump L R, Ridgwell A, et al. 2024. Coupled decline in ocean pH and carbonate saturation during the Palaeocene-Eocene Thermal Maximum. Nat Geosci, 17: 1299–1305

Li X. 2013. Characterization, controlling, and reduction of uncertainties in the modeling and observation of land-surface systems. Sci China Earth Sci, 57: 80–87

Liu Z, Otto-Bliesner B L, He F, et al. 2009. Transient simulation of last deglaciation with a new mechanism for Bolling-Allerod warming. Science, 325: 310–314

Liu Z, Lu Z, Wen X, et al. 2014. Evolution and forcing mechanisms of El Niño over the past 21,000 years. Nature, 515: 550–553

Liu Z, Bao Y, Thompson L G, et al. Tropical mountain ice core $\delta^{18}O$: A Goldilocks indicator for global temperature change. Sci Adv, 2023, 9: eadi6725

Liu Z, Cheng J, Zheng Y, et al. 2025. The seasonal temperature conundrum for the Holocene. Sci Adv, 11: eadt8950

Lyu Z, Vuille M, Goosse H, et al. 2024. South American monsoon intensification during the last millennium driven by joint Pacific and Atlantic forcing. Sci Adv, 10: eado9543

Mann M E, Zhang Z, Rutherford S, et al. 2009. Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly. Science, 326: 1256–1260

Marcott S A, Shakun J D, Clark P U, et al. 2013. A reconstruction of regional and global temperature for the past 11,300 years. Science, 339: 1198–1201

Meng Z, Hakim G J. 2024. Reconstructing the tropical Pacific upper ocean using online data assimilation with a deep learning model. J Adv Model Earth Syst, 16: e2024MS004422

Meng Z, Hakim G J, Steig E J. 2025. Coupled seasonal data assimilation of sea ice, ocean, and atmospheric dynamics over the last millennium. arXiv preprint, arXiv: 2501.14130

Ning L, Liu Z, Mann M E, et al. 2025a. Decadal climate variability during the pre-industrial Common Era: Characteristics and mechanisms. Sci Bull, 70: 2190–2203

Ning L, Xing F, Liu Z, et al. 2025b. Tripolar precipitation change accompanying water isotopes in the Holocene reanalysis of Asian monsoon hydroclimate. Geophys Res Lett, 52: e2025GL116451

Osman M B, Tierney J E, Zhu J, et al. 2021. Globally resolved surface temperatures since the Last Glacial Maximum. Nature, 599: 239–244

Otto-Bliesner B L, Brady E C, Fasullo J, et al. 2016. Climate variability and change since 850 CE: An ensemble approach with the Community Earth System Model (CESM). Bull Amer Meteorol Soc, 97: 735–754

PAGES2k Consortium. 2017. A global multiproxy database for temperature reconstructions of the Common Era. Sci Data, 4: 170088

Pendergrass A G, Hakim G J, Battisti D S, et al. 2012. Coupled air-mixed layer temperature predictability for climate reconstruction. J Clim, 25: 459–472

Peng Z, Lei L, Tan Z M. 2024. A hybrid deep learning and data assimilation method for model error estimation. Sci China Earth Sci, 67: 3655–3670

Perkins W A, Hakim G J. 2017. Reconstructing paleoclimate fields using online data assimilation with a linear inverse model. Clim Past, 13: 421–436

Perkins W A, Hakim G J. 2020. Linear inverse modeling for coupled atmosphere-ocean ensemble climate prediction. J Adv Model Earth Syst, 12: e2019MS001778
Sanchez S C, Hakim G J, Saenger C P. 2021. Climate model teleconnection patterns govern the Niño-3.4 response to early Nineteenth-Century volcanism in coral-based data assimilation reconstructions. J Clim, 34: 1863–1880
Shi F, Goosse H, Klein F, et al. 2019. Monopole mode of precipitation in East Asia modulated by the South China Sea over the last four centuries. Geophys Res Lett, 46: 14713–14722
Snoll B, Ivanovic R F, Valdes P J, et al. 2022. Effect of orographic gravity wave drag on Northern Hemisphere climate in transient simulations of the last deglaciation. Clim Dyn, 59: 2067–2079
Steiger N J, Hakim G J, Steig E J, et al. 2014. Assimilation of time-averaged pseudoproxies for climate reconstruction. J Clim, 27: 426–441
Steiger N J, Smerdon J E, Cook E R, et al. 2018. A reconstruction of global hydroclimate and dynamical variables over the Common Era. Sci Data, 5: 180086
Su J, Chen D, Zheng D, et al. 2023. The insight of why: Causal inference in Earth system science. Sci China Earth Sci, 66: 2169–2186
Sun H, Lei L, Liu Z, et al. 2022. An analog offline EnKF for paleoclimate data assimilation. J Adv Model Earth Syst, 14: e2021MS002674
Sun H, Lei L, Liu Z, et al. 2024. A hybrid gain analog offline EnKF for paleoclimate data assimilation. J Adv Model Earth Syst, 16: e2022MS003414
Sun H, Lei L, Liu Z, et al. 2025. An online paleoclimate data assimilation with a deep learning-based network. J Adv Model Earth Syst, 17: e2024MS004675
Sundqvist H S, Kaufman D S, McKay N P, et al. 2014. Arctic Holocene proxy climate database—New approaches to assessing geochronological accuracy and encoding climate variables. Clim Past, 10: 1605–1631
Talagrand O. 1997. Assimilation of observations, an introduction. J Meteorol Soc Jpn, 75: 191–209
Tardif R, Hakim G J, Perkins W A, et al. 2019. Last millennium Reanalysis with an expanded proxy database and seasonal proxy modeling. Clim Past, 15: 1251–1273
Tian Z, Jiang D, Zhang R, et al. 2020. Transient climate simulations of the Holocene (version 1)—Experimental design and boundary conditions. Geosci Model Dev, 15: 4469–4487
Tierney J E, Zhu J, King J, et al. 2020. Glacial cooling and climate sensitivity revisited. Nature, 584: 569–573
Tierney J E, Zhu J, Li M, et al. 2022. Spatial patterns of climate change across the Paleocene-Eocene Thermal Maximum. Proc Natl Acad Sci USA, 119: e2205326119
Tierney J E, King J, Osman M B, et al. 2025a. Pliocene warmth and patterns of climate change inferred from paleoclimate data assimilation. AGU Adv, 6: e2024AV001356
Tierney J E, Judd E J, Osman M B, et al. 2025b. Advances in paleoclimate data assimilation. Annu Rev Earth Planet Sci, 53: 625–650
von Storch H, Cubasch U, Gonzalez-Rouco J F, et al. 2000. Combining paleoclimatic evidence and GCMs by means of data assimilation through upscaling and nudging (DATUN). In: Proceeding of the 11th Symposium on Global Change Studies. 1: 28–31
Walter R M, Sayani H R, Felis T, et al. 2023. The CoralHydro2k database: A global, actively curated compilation of coral δ^{18}O and Sr/Ca proxy records of tropical ocean hydrology and temperature for the Common Era. Earth Syst Sci Data, 15: 2081–2116
Wan L, Liu J, Gao C, et al. 2020. Study about influence of the Holocene volcanic eruptions on temperature variation trend by simulation (in Chinese). Quat Sci, 40: 1579–1610
Wang H. 2022. How can research on ancient and modern climate and environment be integrated? (in Chinese) Earth Sci, 47: 3811–3812
Wang Z, Liu J, Wang X, et al. 2016. Divergent sensitivity of earth system model CESM 1.0 to solar radiation versus greenhouse gases (in Chinese). Quat Sci, 36: 758–767
Wei Y, Deng W, Chen X, et al. 2024. A comprehensive evaluation of machine learning on coral trace element paleothermometers for sea surface temperature reconstruction. Paleoceanog Paleoclimatol, 39: e2024PA004885
Widmann M, Goosse H, van der Schrier G, et al. 2010. Using data assimilation to study extratropical Northern Hemisphere climate over the last millennium. Clim Past, 6: 627–644
Wu F, Ning L, Liu Z, et al. 2025. A new last two millennium reanalysis based on hybrid gain analog offline EnKF and an expanded proxy database. NPJ Clim Atmos Sci, 8: 62
Xie X, Liu X, Chen G, et al. 2019. A transient modeling study of the latitude dependence of East Asian winter monsoon variations on orbital timescales. Geophys Res Lett, 46: 7565–7573
Yan M, Liu Z, Han J, et al. 2023. Relationship between the East Asian summer and winter monsoons at obliquity time scales. J Clim, 36: 3993–4003
Zhang H, Li M, Hu Y. 2025. Paleoclimate data assimilation: Principles and prospects. Sci China Earth Sci, 68: 407–424
Zhang Q, Berntell E, Axelsson J, et al. 2021. Simulating the mid-Holocene, last interglacial and mid-Pliocene climate with EC-Earth3-LR. Geosci Model Dev, 14: 1147–1169
Zhu F, Emile-Geay J, Hakim G J, et al. 2021. LMR Turbo (LMRt): A lightweight implementation of the LMR framework (0.8.0) [Software]. Zenodo. https://doi.org/10.5281/zenodo.5205223
Zhu F, Emile-Geay J, Anchukaitis K J, et al. 2022. A re-appraisal of the ENSO response to volcanism with paleoclimate data assimilation. Nat Commun, 13: 747
Zhu F, Emile-Geay J, Anchukaitis K J, et al. 2023. A pseudoproxy emulation of the PAGES 2k database using a hierarchy of proxy system models. Sci Data, 10: 624
Zhu F, Emile-Geay J, Hakim G J, et al. 2024. cfr (v2024.1.26): A Python package for climate field reconstruction. Geosci Model Dev, 17: 3409–3431

(Editorial handling: Zhongshi ZHANG)