Temporal Comparisons Involving Paleoclimate Data Assimilation: Challenges & Remedies

Julien Emile-Geay,[a] Gregory J. Hakim,[b] Frederi Viens,[c] Feng Zhu,[d] Daniel E. Amrhein,[d]

[a] *University of Southern California, Department of Earth Sciences, Los Angeles, CA, USA*

[b] *Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA*

[c] *Department of Statistics, Rice University, Houston, TX, USA*

[d] *NSF National Center for Atmospheric Research, Boulder, CO, USA*

Corresponding author: Julien Emile-Geay, julieneg@usc.edu

ABSTRACT: Paleoclimate reconstructions are increasingly central to climate assessments, placing recent and future variability in a broader historical context. Paleoclimate reconstructions are increasingly central to climate assessments, placing recent and future variability in a broader historical context. Several estimation methods produce plumes of climate trajectories that practitioners often want to compare to other reconstruction ensembles, or to deterministic trajectories produced by other means, such as global climate models. Of particular interest are “offline” data assimilation (DA) methods, which have recently been adapted to paleoclimatology. Offline DA lacks an explicit model connecting time instants, so its ensemble members are not true system trajectories. This obscures quantitative comparisons, particularly when considering the ensemble mean in isolation. We propose several resampling methods to introduce a priori constraints on temporal behavior, as well as a general notion, called plume distance, to carry out quantitative comparisons between collections of climate trajectories ("plumes"). The plume distance provides a norm in the same physical units as the variable of interest (e.g. °C for temperature), and lends itself to assessments of statistical significance. We apply these tools to four paleoclimate comparisons: (1) global mean surface temperature (GMST) in the online and offline versions of the Last Millennium Reanalysis (v2.1); (2) GMST from these two ensembles to simulations of the Paleoclimate Model Intercomparison Project past1000 ensemble; (3) LMRv2.1 to the PAGES 2k (2019) ensemble of GMST and (4) northern hemisphere mean surface temperature from LMR v2.1 to the Büntgen et al. (2021) ensemble. Results generally show more compatibility between these ensembles than is visually apparent. The proposed methodology is implemented in an open-source Python package, and we discuss possible applications of the plume distance framework beyond paleoclimatology.

SIGNIFICANCE STATEMENT: Paleoclimate data assimilation is an emerging technique to reconstruct past climate variations. The currently dominant approximation, “offline” data assimilation, lacks the ability to connect information across time. This work proposes open-source solutions to this problem, and applies them to 3 paleoclimate questions, before discussing broader implications.

1. Introduction

In recent years, paleoclimate data assimilation (PDA) has gained traction as a method to estimate variations in past climate fields (Jones and Widmann 2004; Goosse et al. 2006; Gebhardt et al. 2008; Widmann et al. 2010; Goosse et al. 2010; Annan and Hargreaves 2012; Steiger et al. 2014; Hakim et al. 2016; Franke et al. 2017; Acevedo et al. 2017; Steiger et al. 2018; Tierney et al. 2020; Osman et al. 2021; King et al. 2021; Zhu et al. 2022; Shoji et al. 2022; Valler et al. 2022; **?**; **?**). Much like Bayesian hierarchical methods (Tingley and Huybers 2010a,b; Tingley and Huybers 2013), PDA proceeds by drawing from a prior distribution of climate states, which it updates by comparison with observations (Wikle and Berliner 2007). In both cases, the output of these methods is a time-evolving distribution (the “posterior”) quantifying the probability of particular climate states over time. Typically, this (continuous) distribution is discretely sampled and provided in the form of an ensemble, particularly for those DA methods that fall under the general umbrella of Ensemble Kalman Filters [EnKF; Carrassi et al. (2018)].

Summarizing this rich output, for instance to focus on temporal variations, means that such distributions are often reduced to a single representative summary like the mean or median (Büntgen et al. 2020), which in the Gaussian context is the most likely outcome. This presents an apparent paradox: in the parts of the reconstruction least constrained by observations (often, the earliest ones) where the posterior distribution is at its widest (as measured, for instance, by the ensemble variance, or the inter-quartile range), the median often appears very “flat" over time (e.g. see Fig 1a), implying muted variability. Yet, the large spread of this ensemble means that a potentially infinite number of solutions are admitted, some with very high temporal variance, as we will show.

In the Last Millennium Reanalysis (Tardif et al. (2019), Fig. 1) as in many other reconstructions (Steiger et al. 2014; Hakim et al. 2016; Steiger et al. 2018; Neukom et al. 2019; Tierney et al. 2020; Erb et al. 2022; King et al. 2021; Osman et al. 2021; Zhu et al. 2022) this behavior stems from the use of a so-called "offline" DA approach, wherein no explicit rule links different instants in time, so all temporal information is provided by the paleoclimate proxy data (for more details, see Sect. 2). Where this information is dense and reliable, the posterior distribution is relatively tight, and the temporal behavior of the median/mean well-constrained. Where this information is sparse and/or noisy, the posterior distribution is spread out, and the temporal behavior of the median (Fig. 1, gold line) or any random path (Fig. 1a, orange and blue lines) are relatively flat. This is not an issue if the full ensemble, or a meaningful summary of its spread (Fig. 1a-c), are provided to users; however, in many applications, only the mean or median is provided. This narrow focus can lead to the misleading impression that reconstructed climate trajectories lack temporal variability (Neukom et al. 2022), or that several competing series (e.g. reconstructions or model simulations) are less compatible with the DA ensemble than they really are. For instance, Fig. 1b shows how this ensemble fares compared to simulations from the Paleoclimate Model Intercomparison Project (PMIP) 3 (Dufresne et al. 2013; Giorgetta et al. 2013; Gordon et al. 2000; Otto-Bliesner et al. 2015; Rotstayn et al. 2012; Schmidt et al. 2012, 2006; Stevenson et al. 2019; Watanabe et al. 2011; Wu et al. 2014), while Fig. 1c compares LMRv2.1's reconstructed Northern Hemisphere temperature to the median reconstruction of the same quantity from Büntgen et al. (2021). Such representations allow qualitative comparisons, but raise the question of how to quantify the compatibility between such traces[1] and an offline DA ensemble like LMRv2.1.

In light of the growing use of offline DA ensembles in climate studies (Singh et al. 2018; Erb et al. 2020; Zhu et al. 2020; Tejedor et al. 2021; Osman et al. 2021; King et al. 2021; Zhu et al. 2022; Dee and Steiger 2022; Erb et al. 2022), it appears timely to clarify what information may be derived from such offline DA ensembles, what information may be lost in the reconstruction process, and what post-hoc adjustments may be performed to remedy the situation. In this paper we discuss the interpretation and use of such ensembles for various applications, and introduce open-source tools that can be used to estimate temporal properties of these data products under fairly strong assumptions. To simplify the exposition, we focus on summary scalar measures like

[1] A timeseries $y(t)$ is often called a "trace"; in the following, we use these terms interchangeably.

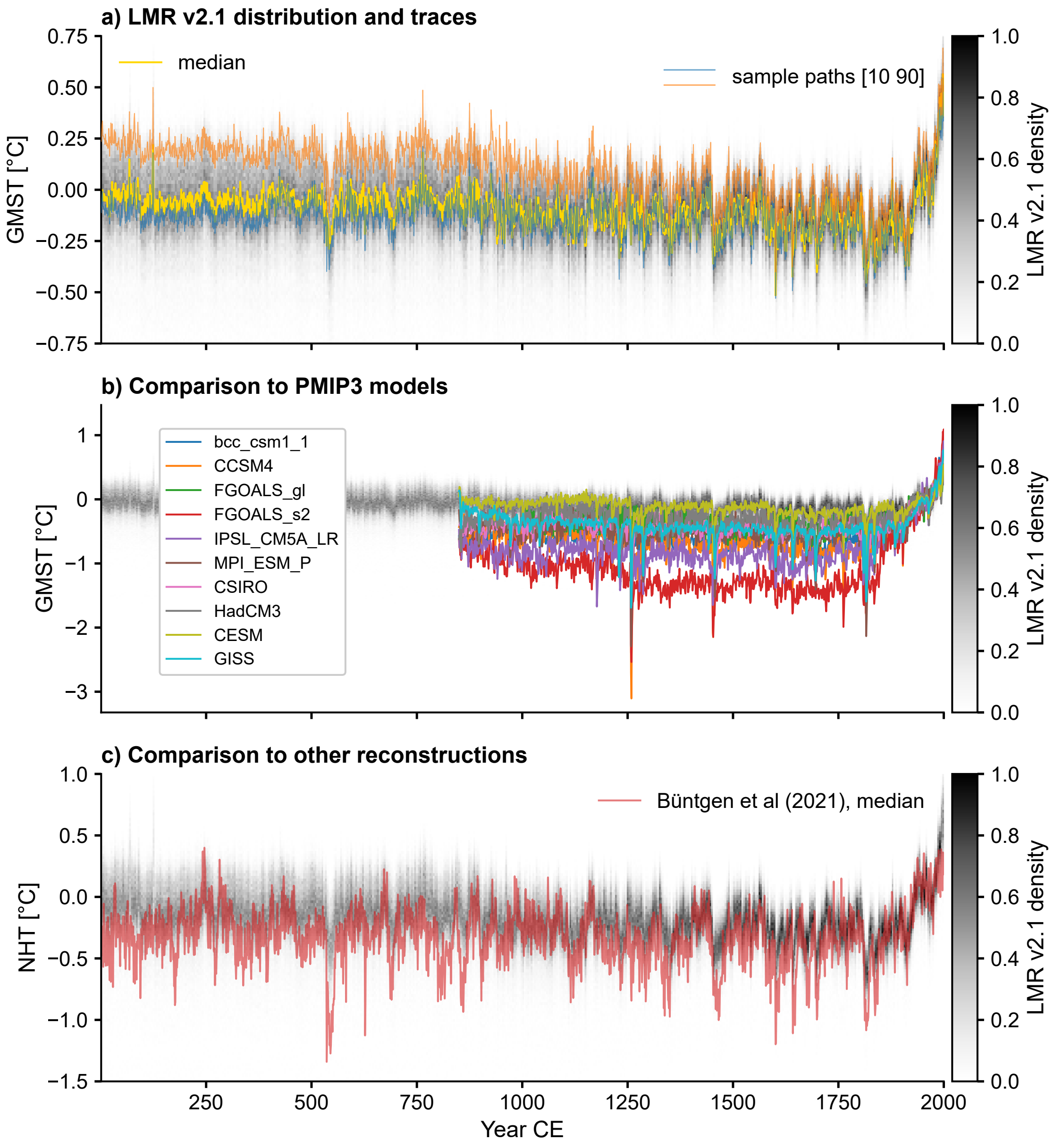

FIG. 1. **The LMRv2.1 global mean surface temperature (GMST) and some comparisons of interest**. All three panels show the posterior density of GMST shaded in gray. In a), the colored lines represent 2 sample paths through the ensemble, labeled arbitrarily (see Sect. 2). b) Comparison of the LMRv2.1 GMST posterior density to `past1000` simulations of the Paleoclimate Model Intercomparison Project (PMIP) 3 (Dufresne et al. 2013; Giorgetta et al. 2013; Gordon et al. 2000; Otto-Bliesner et al. 2015; Rotstayn et al. 2012; Schmidt et al. 2012, 2006; Stevenson et al. 2019; Watanabe et al. 2011; Wu et al. 2014). c) same as b), comparing the LMRv2.1 assimilated Northern Hemisphere Temperature to the median reconstruction of the same quantity from Büntgen et al. (2021).

global or hemispheric mean surface temperatures, leaving the treatment of the full spatial problem for future work.

We start with a brief recount of the properties of offline DA (Sect. 2), before assessing similarity within an offline DA ensemble (Sect. 3). This leads us to parametric modeling choices that can best preserve temporal structure. We show that notions of proximity or likelihood in such a space are non-trivial and motivate the introduction of a new pathwise measure, called proximity probability, from which a distance metric can be derived (Sect. 4). We then apply these concepts to comparing reconstructions of global mean surface temperature, and comparing reconstruction ensembles to climate simulations (Sect. 5). Discussion follows in Sect. 6. Technical details are provided in the appendices.

2. Offline Data Assimilation

Offline DA stands in contrast with "online" DA methods (used for instance in numerical weather prediction and more rarely in paleoclimate reconstructions (Widmann et al. 2010; Franke et al. 2017; Perkins and Hakim 2017; Amrhein et al. 2018; Perkins and Hakim 2021)), wherein a physically-based model is used to propagate climate states through time. Online DA methods explicitly model the system's temporal evolution, and are as such more desirable, yet often more costly to implement. In cases where the predictive skill of a given model is marginal, offline DA provides a competitive solution, trading off computational expediency for a lack of explicit temporal constraints.

Given the importance of these reconstructions in providing historical context for recent warming trends (IPCC 2021, Fig 1), it is critical to account for the uncertainty in these reconstructions when, for example, testing hypotheses. These ensemble methods sample from a posterior distribution of climate states involving a weighting of information from observations (proxies) and model prior. The individual ensemble members are equally likely, so any trajectory encompassed by these distributions is technically allowed, which creates challenges for comparing the temporal behavior of reconstructions with each other, and reconstructions with models.

While the ensemble time series for time-integrated methods, such as from a climate model or online data assimilation, are distinct, the ensemble members for offline data assimilation have no temporal linkage. For offline data assimilation, there is no forecast step linking assimilation times, and time-independent ensembles (i.e. fixed collections of climate states) are typically used as the prior at each assimilation time. In order to discuss the consequences of this common approximation in posterior analyses involving time, we first briefly review the Kalman filter.

Given a prior estimate of the climate state, with mean $\mathbf{x}^b$ and error covariance matrix $\mathbf{P}^b$, at a time for which we have observations in the form of paleoclimate proxies, $\mathbf{y}$, with error covariance matrix $\mathbf{R}$, the minimum variance estimate of the true state mean is given by

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b) \tag{1}$$

with error covariance

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^b. \tag{2}$$

Here, $\mathbf{H}$ maps from the climate state to the observations (proxies). The weight given to the novel information from observations is determined by the Kalman gain matrix

$$\mathbf{K} = \mathbf{P}^b\mathbf{H}^\mathrm{T}(\mathbf{H}\mathbf{P}^b\mathbf{H}^\mathrm{T} + \mathbf{R})^{-1}. \tag{3}$$

Offline DA methods approximate solutions to (1) and (2) using ensembles that are typically drawn from existing long climate model simulations, the details of which are not important here. The key is that the same sample is used to estimate the climate statistics at each time, so that the estimate of $\mathbf{P}^b$ is independent of time. While different samples can be drawn for each time, the resulting $\mathbf{P}^b$ differ only by sampling error, not due to physics (that is, these errors are uncorrelated in time, within sampling error). As a consequence, the only time variation in $\mathbf{K}$, and hence $\mathbf{P}^a$, comes from time variation in the availability of observations. In the limit of a fixed observing network, $\mathbf{K}$ and $\mathbf{P}^a$ are constant in time; the ensemble perturbations that sample $\mathbf{P}^a$ are therefore also constant in time. Time series for the i-th ensemble member, for any scalar, such as one grid point for one variable, can be expressed as a sum of the ensemble mean x^a, derived from (1), and the ensemble perturbation x_i', derived from (2):

$$x_i^a = x^a + x_i'. \tag{4}$$

By construction, x_i' has zero mean and covariance $\mathbf{P}^a$. Thus, while x^a depends on time through the observed values, $\mathbf{y}$, each perturbation x_i' depends only on the time-availability of the observations

(**H**) and their errors (**R**) – see (2). Since the label i is arbitrary, it may be changed without any effect on the estimates for $\mathbf{x}^a$ and $\mathbf{P}^a$. Error estimates for any inference or calculation involving the ensemble as a function of time must consider the freedom to relabel the ensemble members in time, which generates new time series.

Here we consider the impact of this lack of temporal constraint on the ensemble members. We begin with comparisons between the ensemble mean and individual members with other deterministic time series, which highlights signal vs. noise problems. We then show how nonlinear temporal measures, like power spectra, are affected by uncorrelated errors. With that motivation we then propose several approaches to introduce physically-realistic temporal dependence to the offline ensembles and show the impact on various diagnostics, both linear and nonlinear.

3. Ensemble neighbors

A common paleoclimate question may be phrased thus: how compatible is a given reconstruction with another, or with a model simulation? Such a question underlies popular summaries like Fig 6.10 from IPCC (2007) or Fig 5.7 from Masson-Delmotte et al. (2013). Consider for instance the simulation of GMST by the HadCM3 (Gordon et al. 2000; Pope et al. 2000) `past1000` simulation from the Paleoclimate Model Intercomparison Project, version 3 (Braconnot et al. 2012). Its trace is plotted in Fig. 1b, along with other last millennium simulations, where they may be compared with the LMRv2.1 posterior density (grayscale). While this visualization allows for a qualitative assessment of similarity, a more precise question is to ask if a close match can be found within the ensemble. That is: can the LMR ensemble be mined for a trace that approximates a target such as the HadCM3 GMST as closely as possible? We call such traces "ensemble neighbors", or simply "neighbors".

a. Naïve Resampling

The simplest approach to finding such neighbors is to minimize the mean squared error between the trace and the ensemble, an approach we call "naïve resampling" because it is oblivious to the implications of the resampling for temporal variability, which will be apparent shortly. Under such a naïve scheme, it is indeed possible to find a very close match (Fig. 2a), which correlates with the target above 0.99. Thus, despite the apparent discrepancies of Fig. 1b, one would conclude that the HadCM3 trace is highly compatible with the LMR ensemble. Repeating this exercise with the other simulations featured in Fig. 1b, a LMR path correlating with each trace above 0.97 can always be found. This is also the case with the red trace in Fig. 1c, and with the 15 reconstructions of northern hemisphere summer temperature on which it is based (not shown).

While a close match may be found in all these cases, this is only possible because of the atemporality of offline DA, where ensemble members are arbitrarily labeled (Sect. 2). This raises two key questions:

Temporal structure: what are the temporal consequences of drawing at random from the ensemble's posterior distribution? How does it affect the ensemble's temporal behavior, and is this physically defensible?

Likelihood: how likely is a given neighbor in the context of the ensemble? In other words, how far into the tails of the ensemble's distribution must the samples be drawn to find the closest match? If the neighbors are only found in the most extreme quantiles of the ensemble, how compatible is the target with the ensemble?

Mining the posterior distribution for values that closely match a target (Fig. 2a) implicitly assumes that all values are equally plausible. This has drastic consequences for estimated variability: Fig. 2b shows the LMR v2.1 ensemble (median and 95% highest density interval)[2] as well as 3 traces obtained by drawing uniformly at random from the posterior at each time step (naïve resampling), resulting in much more erratic trajectories. The frequency-domain consequences of this resampling are shown in the bottom row of Fig. 2: panel c shows the spectral density of the original LMRv2.1 GMST ensemble (red) as well as the spectral density of the ensemble median (blue). In this instance, the median of the ensemble of spectra closely resembles the spectrum of the ensemble's median timeseries; both show near fractal scaling with an exponent $\beta \simeq 1.04$, consistent with previous work (Zhu et al. 2019). This stands in sharp contrast to the spectra of the resampled ensemble (panel d): because of the uniform resampling, the spectra are whitened, with an average spectral slope close to 0.76 (not shown). While the ensemble median (blue curve) is unaffected by resampling, the individual paths very much are, and so is the distribution of spectra (red). This whitening contradicts the near-fractal scaling behavior known to characterize GMST variability

[2] The highest density interval (HDI), or highest density region (HDR), is defined as the most compact region containing a given mass of the distribution, say 95%. In simple cases, this coincides with the 2.5%-97.5% quantiles of a distribution, but is a more general notion. For a more precise definition, see Hyndman (1996).

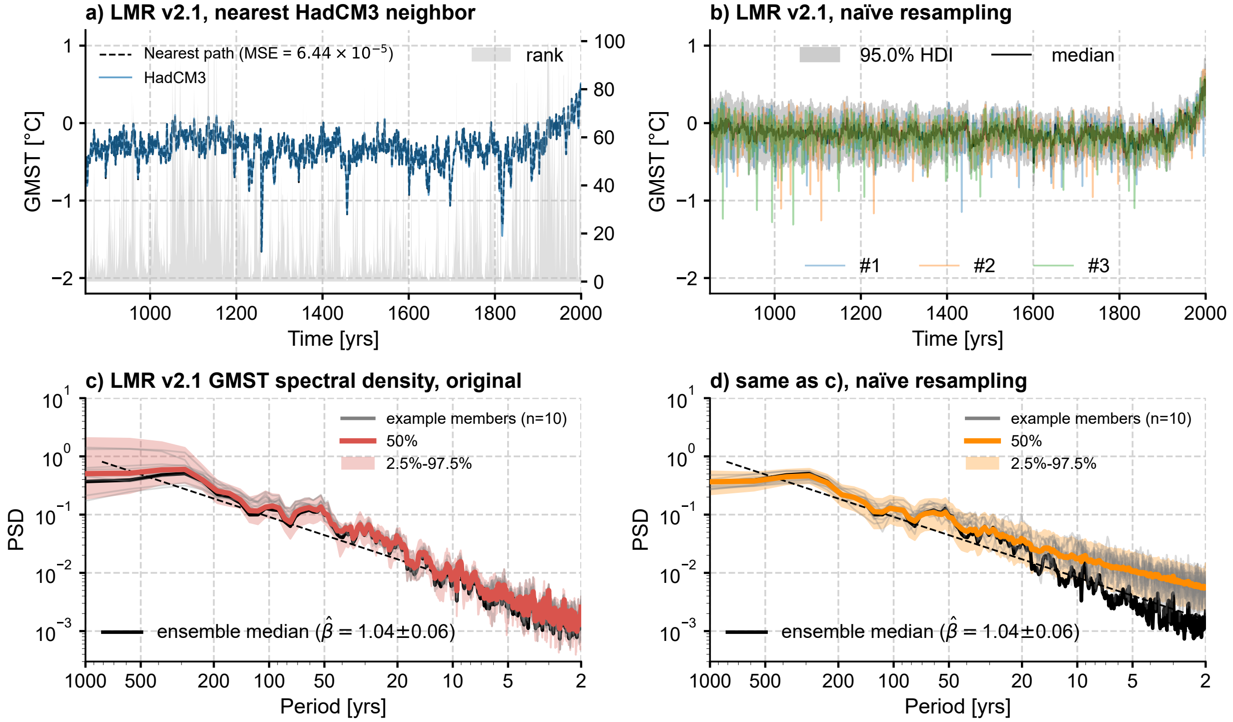

Fig. 2. **Effects of atemporality**. a) The HadCM3 simulation's GMST trace (blue) and its closest neighbor in the LMR v2.1 ensemble (dashed black), obtained by naïve resampling. The gray curve displays the rank of the ensemble members (as percentiles) that were picked to match the HadCM3 trace in each year; notice how ranks are concentrated in the bottom half, and sometimes the very lowest ranks of, the LMR ensemble. b) LMR ensemble along with three random paths obtained by naïve resampling, to illustrate the temporal implications of mining the ensemble for neighbors. c, d) Multitaper GMST spectra (Thomson 1982) of the LMR v2.1 ensemble, computed using Pyleoclim (Khider et al. 2022) with an anti-alias filter (Kirchner 2005). In (c) the spectra come from the original offline DA ensemble (red), with 10 random draws shown in gray. The spectrum of the ensemble median is shown in black, and roughly coincides with the median of the distribution of spectra (thick red curve). Panel (d) shows the same quantities, but for the LMRv2.1 ensemble processed with naïve (uniform) temporal resampling at each time step (as in b). Individual ensemble members show greater variability and a whiter spectrum, but the spectrum of the ensemble median (black) is nearly unchanged, with identical scaling exponents ($\hat{\beta}$) within uncertainties.

over the instrumental era (Fraedrich et al. 2004; Huybers and Curry 2006; Laepple and Huybers 2014; Lovejoy 2015; Fredriksen and Rypdal 2016; Franzke et al. 2020; Hébert et al. 2022), and a reconstruction of the past millennium obtained using online DA (Perkins and Hakim 2021). The latter is shown in Fig. 3, and provides an important cross-check on the offline DA solution. Unlike the latter, this online DA estimate explicitly links climate states through time, using a first-order propagator (a linear inverse model, or LIM (Perkins and Hakim 2017, 2020)). As a result, each individual path through the ensemble (colored traces in Fig. 3a) exhibits a more stable and realistic temporal variability; this variability is also similar to the median's. Both of these characteristics differ markedly from the offline DA solution (Fig. 1a). In the frequency domain, each online DA solution exhibits near-fractal scaling (linear behavior with slope near unity in Fig. 3b's log-log representation), with a sharply peaked distribution of exponents (Fig. 3c). The ensemble median exhibits a very similar exponent of 1.08 ± 0.07, very near the mode of the distribution of individual traces (Fig. 3b).

So while it is possible to pick any trajectory within an offline DA ensemble, it is paramount for this choice to respect the known temporal characteristics of the underlying climate signal. As we have shown, neither the original traces (Fig. 1a) nor their counterparts obtained by naïve resampling (Fig. 2b) achieve this. One must therefore construct sampling rules for the offline DA ensemble that obey independent constraints about climate variability.

b. Parametric Resampling

In the particular case of LMRv2.1, a reconstruction using the same input data and an online DA algorithm are available (Perkins and Hakim 2021), and may be used to provide guidance. In general, this will not be the case, yet there always exist prior constraints on the temporal variability of the target state variable. For instance, theoretical models (inspired by observations) may guide the choice of a random walk (Hasselman 1976) or scaling behavior (Lovejoy and Schertzer 2013; Franzke et al. 2020). This intuition may also come from independent instrumental or proxy observations (Huybers and Curry 2006; Zhu et al. 2019) or from general circulation models, though the latter are known to harbor regional and local biases (Laepple and Huybers 2014; Laepple et al. 2023). One way or another, something is known about the expected temporal structure of the fluctuations, even if only in a gross sense.

Since much existing theory applies to processes with zero mean and unit standard deviation, we first consider the spectral behavior of fluctuations around the ensemble mean: the bottom row of

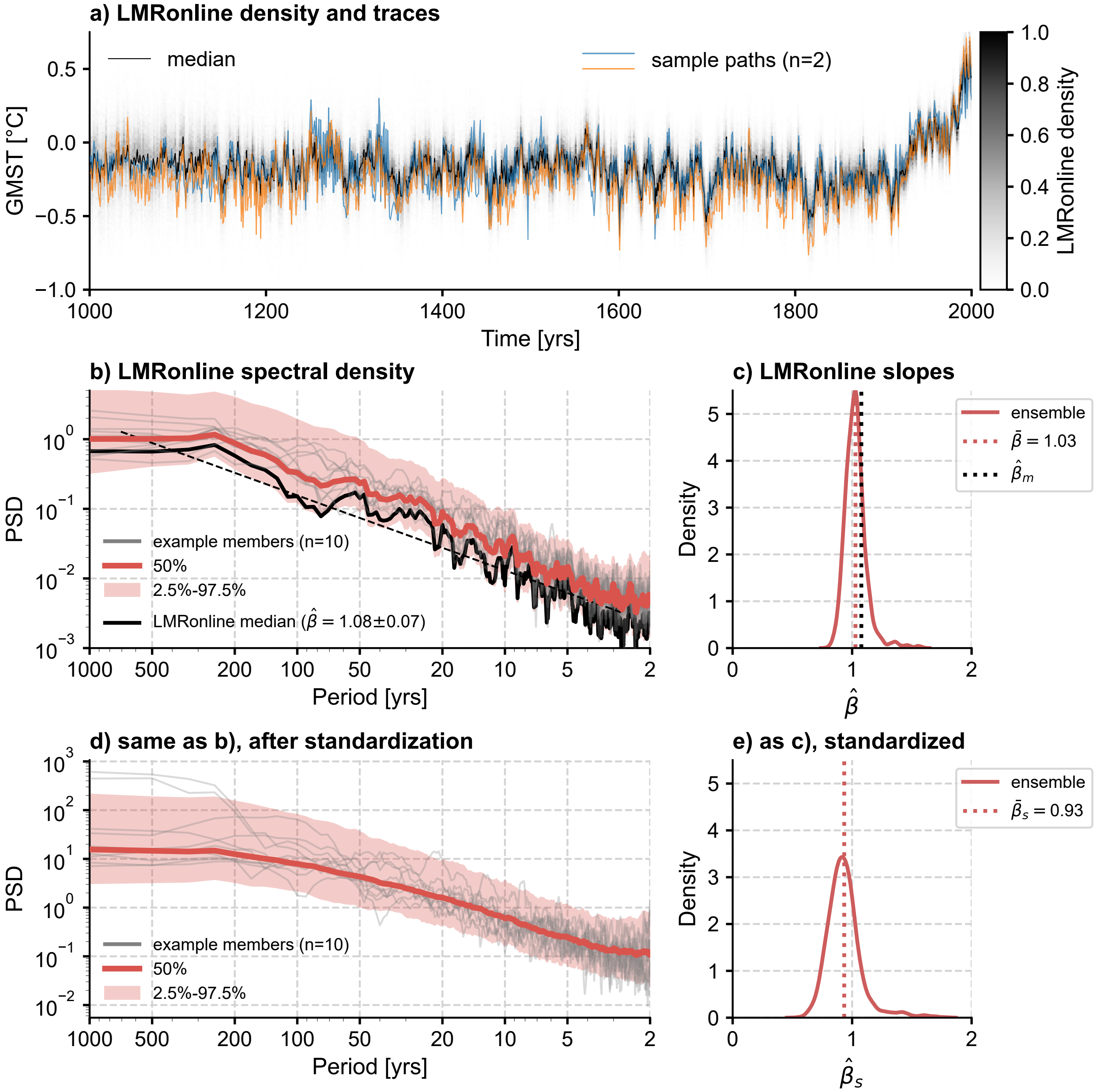

FIG. 3. **The LMRonline GMST**. As Fig. 1a, but for the LMRonline reconstruction of Perkins and Hakim (2021). Notice how each ensemble trace shows a similar level of variability to the ensemble median, unlike the offline ensemble. b) Spectral density of the ensemble shown in a); c) distribution of scaling exponents of the spectra shown in b). d) same as b), after removing the ensemble mean and dividing all traces by the ensemble standard deviation. The spectral density of the ensemble median is omitted as that series is close to 0 at all times, by construction. e) distribution of scaling exponents of the spectra shown in d)

Fig. 3 provides evidence compatible with scaling behavior with slightly flatter slopes than the full signal (Fig. 3e).

The standardized fluctuations are compatible with a power-law spectrum with $\beta \approx 0.93$, though this is not the only possible model fit. Indeed, it is known that long-range dependence can be difficult to distinguish from the superposition of short-range dependencies with different timescales (Maraun et al. 2004), which would be better captured by an autoregressive process. Accordingly, the standardized LMRonline ensemble of (Fig. 3d) can be fit quite closely using an autoregressive model of order 2 (Fig. 4), whose residuals are Gaussian, unstructured, and uncorrelated in time (not shown), indicating a good fit.

The larger point is that there is no unambiguous choice of model to describe GMST fluctuations over the Common Era. Given the behavior observed in Fig. 3 (c,e), we propose 3 models to characterize reconstructions of GMST fluctuations around the ensemble mean obtained via offline DA:

1. an autoregressive model of order p, or AR(p).

2. fractional Gaussian noise (fGn)

3. power-law spectra

Details on the models and their mathematical formulation are given in Appendix 6. Because empirical evidence can be found to support any of those models for GMST fluctuations, we refrain from imposing this choice on users of this framework. Instead, we designed a flexible resampling interface that allows users to specify any of these models, and we encourage more to be added if appropriate.

Fig. 5 shows the result of resampling the LMR v2.1 output according to these three models, using parameters meant to approximate the behavior of the LMRonline solution. Because each of these models assumes stationary noise increments, each trajectory must be scaled so that the ensemble variance $\sigma(t)$ matches that of the original offline DA solution, with uncertainties growing back in time (e.g. Fig. 1a). The ensemble mean is preserved as well, by construction. Therefore, this resampling leaves the ensemble statistics unchanged, but changes the temporal statistics of individual trajectories, which affects comparisons to other reconstructions and model simulations.

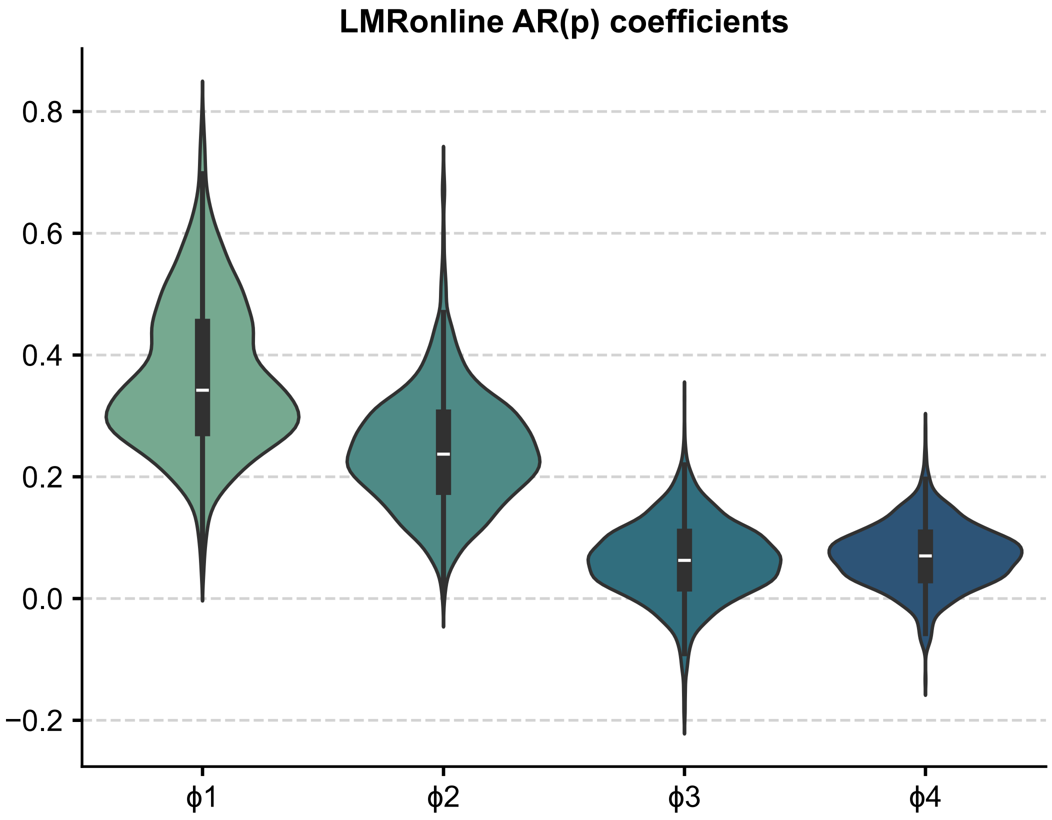

FIG. 4. **Autoregressive modeling of LMRonline trajectories**. Violin plots of the distribution of the first four AR coefficients ($\phi_1, \cdots, \phi_4$) for 2000 randomly-drawn ensemble members from the LMRonline GMST reconstruction of Perkins and Hakim (2021). The model is fit on deviations around the ensemble mean, whose spectrum is shown in Fig. 3d. Each "violin" shows the shape of the distribution, as well as box & whiskers plot showing the inter-quartile region. Median values are shown as white notches. Coefficients beyond order 2 are not meaningfully different from 0, indicating that $p = 2$ is an appropriate choice of order for this fit.

4. Assessing ensemble proximity

We now return to the question of proximity raised in Sect. 3.a: how likely is a given trace in the context of an ensemble? Consider the case presented in Fig. 6, where one wishes to compare two traces $y_1(t)$ and $y_2(t)$ to an ensemble of trajectories $X_i(t)$, where t indexes time and $i \in \mathbb{N}$ indexes ensemble members. Visually, it is obvious that the HadCM3 trace is more closely compatible with

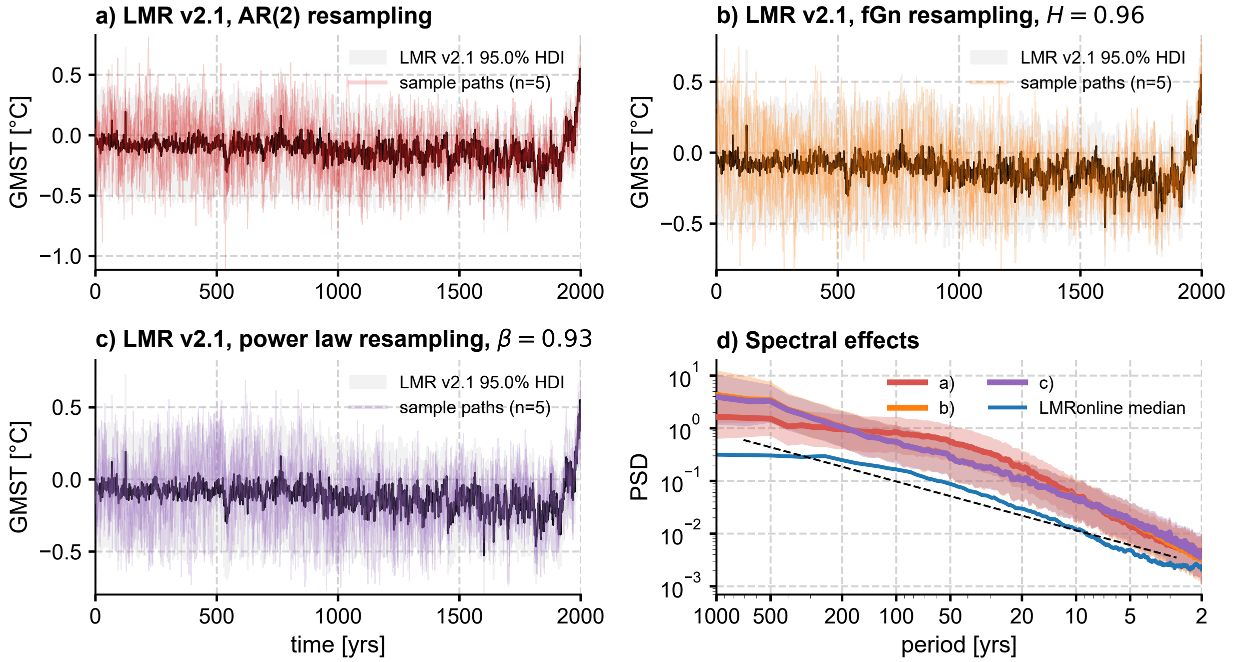

FIG. 5. **Parametric Resampling of offline DA output**. a- c): gray envelopes show the 95% Highest Density Interval (HDI) from the LMR v2.1 output, with the ensemble mean in dark gray. Thin, colored lines show the temporal evolution of 10 randomly-drawn traces under the three models considered in the text: a) AR(2), b) fractional Gaussian noise, and c) power-law scaling (see text for details). Panel (d) shows the spectra of these solutions, and how they approximate the spectrum of the LMRonline solution of Perkins and Hakim (2021), unlike naïve resampling (Fig. 2d).

LMRonline than CCSM4, and here we explore a new method to quantify time series similarity to an ensemble.

a. Proximity Probability and Plume Distance

A natural approach to similarity assesses the likelihood of each trace given the ensemble X_i from which it is drawn, and compute the likelihood ratio between them. However, the high-dimensionality of the sample space ($T = 2001$ time points), typically leads to vanishingly small numbers for the likelihood of a given trace (see Appendix B). While there exist many mathematical tools to quantify the compatibility of a point with an ensemble (e.g. from the forecast verification literature (Gneiting and Katzfuss 2014)), these tools are not well suited to our particular problem: quantifying similarity between a trajectory, or ensemble of trajectories, to a time-evolving distri-

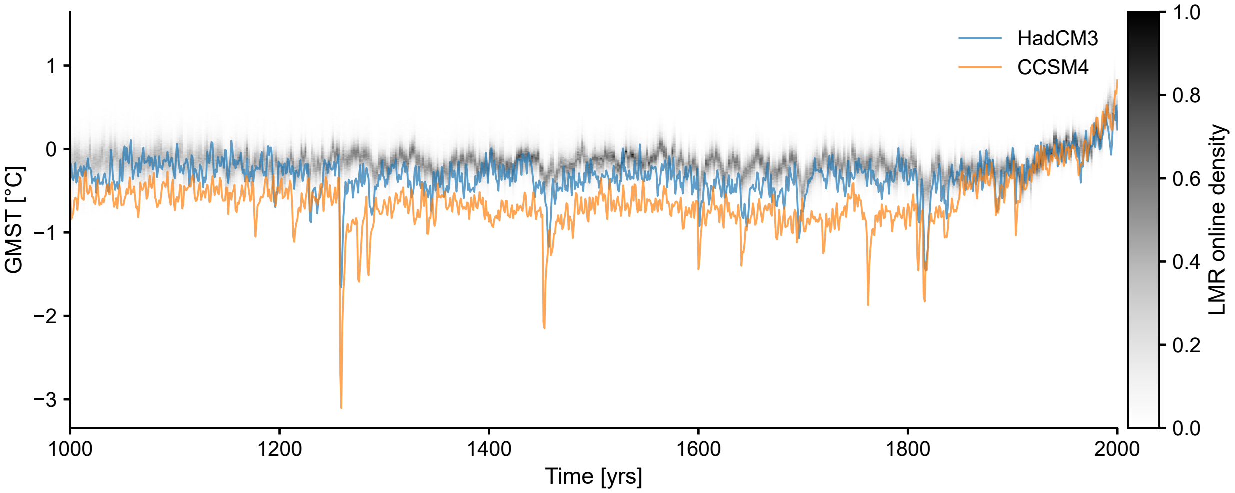

FIG. 6. **Proximity to an ensemble.** Shown here are the HadCM3 and CCSM4 `past1000` GMST (colored traces) described in Fig. 1b. They are compared to the LMRonline (Perkins and Hakim 2021) posterior density.

bution. Our problem is related to the "shadowing trajectory" challenge for dynamical systems, and difficulties in using observations to distinguish trajectories in high-dimensional systems (Judd and Smith 2004; Judd et al. 2008).

We introduce a proximity metric that uses a finite scale of comparison, instead of infinitesimal volumes implicit in the use of probability densities and similar likelihood concepts for high-dimensional or continuous state-space settings. Further theoretical justification for this metric may be found in Appendix C. Our approach is as follows: given an ensemble $X_i(t), i \in [1, \cdots, p]$ and a trace $y(t)$, consider a tube around $y(t)$ of size ϵ, and shape determined by a norm on trace space, such as the ℓ^q norm, for some number $q \in [1, \infty]$. One then enumerates the number of ensemble trajectories $i = 1, 2, \cdots, p$ that fit entirely within that tube. Specifically, the procedure is as follows:

1. compute the q-norm distance between a trace y and each of the p ensemble members.

2. graph the distribution of distances $d_q(y, X) = \|y - X\|_q$, as X ranges over all p ensemble members, to choose a sensible range of ϵ parameters (e.g. Fig. 7a).

3. Compute the (empirical) proximity probability $\mathbb{P}(d_q(y, X) \leq \epsilon)$ as the proportion of ensemble members that fit within the tube for a given set of ϵ parameters.

4. Graph this proportion as a function of ϵ (Fig. 7b).

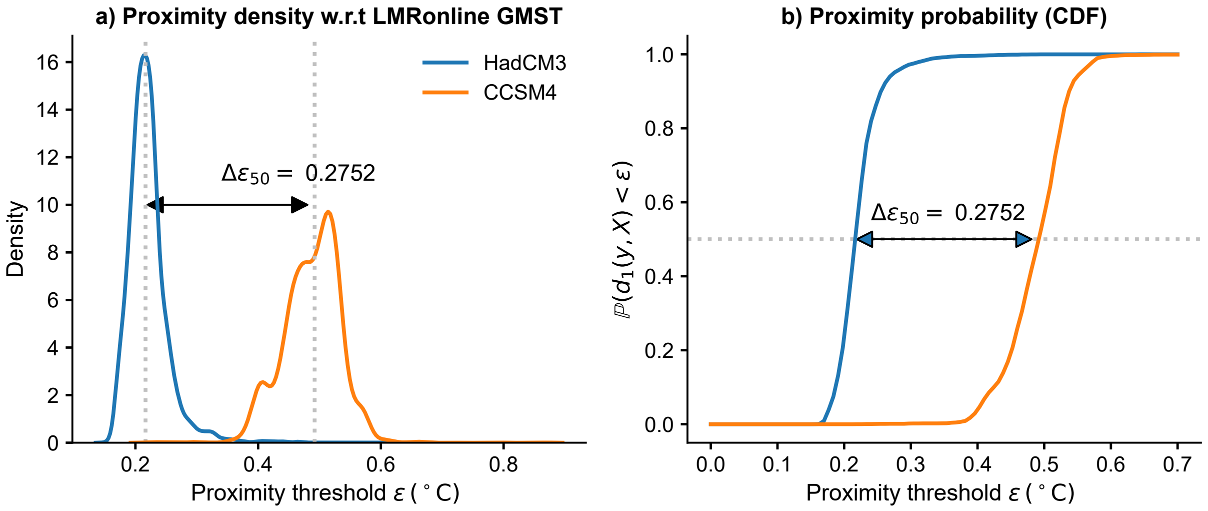

FIG. 7. **Proximity Statistics**, including the kernel density estimate of distances between the HadCM3 and CCSM4 past1000 traces and the LMRonline ensemble (left), as well as the cumulative density function based on those distances (right) which we term "proximity probability". The arrow illustrates the "plume distance" concept, evaluated in this case to be approximately 0.28°C. The norm used here is the ℓ^1-norm (d_1).

In effect, each of these graphs of proportions as a function of ϵ is the empirical cumulative distribution function (CDF) of the distance from the fixed trace y to the ensemble viewed as a random trajectory X (Appendix C). These "proximity probabilities" can be leveraged to compute simple, robust statistics of distance. Any non-tail percentile of the proximity probability, which is measured in the same units as y or X (here, °C of GMST), may be used for this purpose. Fig. 7 illustrates this metric for the 50% quantile, though it is nearly unchanged anywhere between the 20% and 80% quantiles. Remarkably, the metric is also extremely stable to the choice of norm ($q = 1, 2, \infty$), varying only within 10^{-3} in this example (not shown).

We propose the proximity probability for the 50% quantile, which we call the *plume distance*, as a useful and robust summary of the distance between an ensemble (plume of trajectories) and a target (Appendix C). In this case, it says that the HadCM3 trace is closer to the LMRonline ensemble than the CCSM4 trace by about 0.28°C. However, like all summary statistics, it results in a loss of information. To report a fuller assessment of the uncertainty profile for the distance from the ensemble to the target, one may also graph the proximity probability (Fig. 7b) or its derivative, the *proximity density* (Fig. 7b).

In the following sections we show how to use these measures in various comparisons. One notion left to be worked out is that of significance: if a plume distance of 0.5°Cis found between two ensembles, or between a trace and an ensemble, it is natural to ask how significant this distance is compared to the inherent spread of the ensemble used as benchmark for the comparison. We explore this question using a comparison between the offline and online versions of LMR.

b. Intra- vs inter-ensemble distances

Perkins and Hakim (2021) compared their reconstruction ("LMRonline") to the offline DA version LMR v2.1 (Tardif et al. 2019), and found that the LMRonline median exhibited larger temporal variability, and its distribution was much tighter (smaller HDI), than LMR v2.1. Still, it is worth asking whether these two products, based on the same inputs (proxy data, model prior), are compatible by our proximity metric. Two key notions here are those of **inter-ensemble distances** (distances between pairs of trajectories from each ensemble, for a given set of proximity thresholds ϵ) and **intra-ensemble distances** (distances between pairs of trajectories within an ensemble, for a given set of proximity thresholds ϵ). The plume distance defined above is merely the median of the distribution of inter-ensemble distances.

Fig. 8 (left) compares the plume distance between those two ensembles with the distributions of their intra-ensemble distances. Because the LMRonline ensemble is denser than LMRv2.1 (5000 vs 2000 members), we first cull it by selecting 2000 trajectories at random, to ensure a meaningful comparison; results shown here are insensitive to the stochastic realization of this selection. The ℓ^1 norm was used, though results are also insensitive to this choice.

Fig. 8 (left) shows that the plume distance ($\Delta\epsilon_{50}$) coincides approximately with the mode of the LMRv2.1 proximity density. The LMRonline distances are clustered relatively tightly around 0.12, and are entirely encompassed by the much wider range of distances found amongst LMR2.1 traces. This suggests that these two ensembles are compatible with each other: the typical distance between ensembles (i.e., the plume distance, 0.14°C) is entirely within the range of intra-ensemble distances.

Is this result an artifact of the lack of temporal variability in individual traces in the LMRv2.1 ensemble (cf Fig. 1a)? To be sure, we resampled the LMRv2.1 ensemble according to a power-law model with $\beta = 0.93$, as this model is a fair approximation of the actual spectrum (Fig. 3e). The

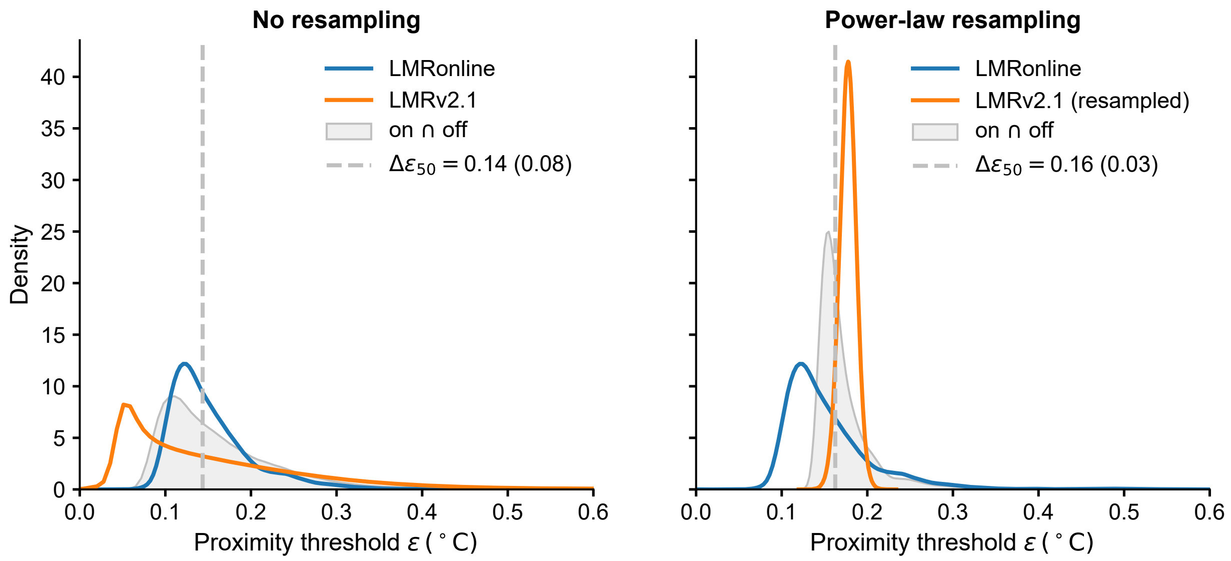

FIG. 8. **Inter- and intra-ensemble distances** for two LMR ensembles: LMRonline (Perkins and Hakim 2021) and LMR v2.1 (Tardif et al. 2019). (left) original LMRv2.1 ensemble; (right) LMRv2.1 ensemble resampled according to a power-law model with $\beta = 0.93$. The variable assessed is GMST in °C, and the blue distribution is common to both plots. The inter-ensemble proximity density is shown in light gray. Its median (the plume distance $\Delta\epsilon_{50}$) is indicated by the dotted gray lines, and is nearly identical between the two cases, but the width of the distribution varies greatly: an interquartile range of 0.08 without resampling, compared to 0.03 with resampling – as reflected by the tighter distribution.

result (Fig. 8, right) shows that resampling has a profound effect on the width of the intra-ensemble distribution (orange), but in this instance the plume distance is nearly unchanged under resampling. Instead, its precision (as measured by the interquartile range of the inter-ensemble distribution) goes from 0.08 (without resampling) to 0.03 (with resampling). Now the roles are reversed: the LMRv2.1 distribution sits within that of the LMRonline ensemble, and the updated plume distance (0.16) appears typical of LMRonline intra-ensemble distances, coinciding nearly perfectly with the mode of its distribution. Again, we conclude that the ensembles are compatible, since one can fit within the other according to our distance metric.

The intra-ensemble distribution also provides a sensible null against which to judge the significance of the plume distance. For instance, one may declare that a trace (or ensemble) is incompatible with a given offline DA ensemble if the plume distance to this ensemble exceeds the 95th percentile of its intra-ensemble proximity density. Alternatively, one may count the fraction of trajectories that lie beyond such a quantile. As always, it is worth emphasizing that the 95th percentile is an arbitrary threshold, and it may be adjusted according to a user's needs or confidence/credibility preferences.

To summarize, Fig. 9 illustrates our process of plume-to-plume comparison with two LMR ensembles: LMR v2.1 (Tardif et al. 2019) and LMRonline (Perkins and Hakim 2021), shifted downward by 0.75°C for illustrative purposes. The plume distance is the median of the distribution of inter-ensemble distances, obtained by randomly selecting traces, drawing tubes of width ϵ around them, and counting how many traces from the other ensemble fit within this tube. Importantly, the plume distance applies equally to comparing an ensemble to a trace or comparing two ensembles; this generality is an appealing aspect of our framework.

5. Applications

We now apply this framework to three paleoclimate comparisons: comparing model simulations to the Last Millennium Reanalysis (Section 5a); comparing results from a multi-method ensemble including offline DA (Section 5b), and comparing the LMRv2.1 ensemble to a heterogeneous ensemble of reconstructions (Section 5c). Each of these examples illustrates different aspects of our methodology.

a. Data-model comparisons over the past millennium

Intra-ensemble distances are natural points of comparison to establish the significance of a plume distance. We apply this logic to an assessment of compatibility between LMRv2.1 GMST and the `past1000` PMIP3 simulations of Fig. 1b. As before, we use the LMRv2.1 GMST ensemble resampled to mimic the LMRonline GMST spectrum (Fig. 3b), according to the three parametric models of Sect. b.

Because 40 comparisons are carried out (10 models, 4 ensembles), it is useful to summarize them via the plume distance ($\Delta\epsilon_{50}$) introduced earlier. This is done in Table 1, where it can be seen that, with a 95% quantile threshold, the LMRonline plume is compatible with 6 simulations (FGOAL_gl, MPI_ESM_P, CSIRO, HadCM3, CESM and GISS), while the (resampled) LMRv2.1 plumes (regardless of the resampling scheme) are only compatible with the CESM simulation. This discrepancy arises for two reasons: 1) the LMRonline intra-ensemble distribution is more diffuse

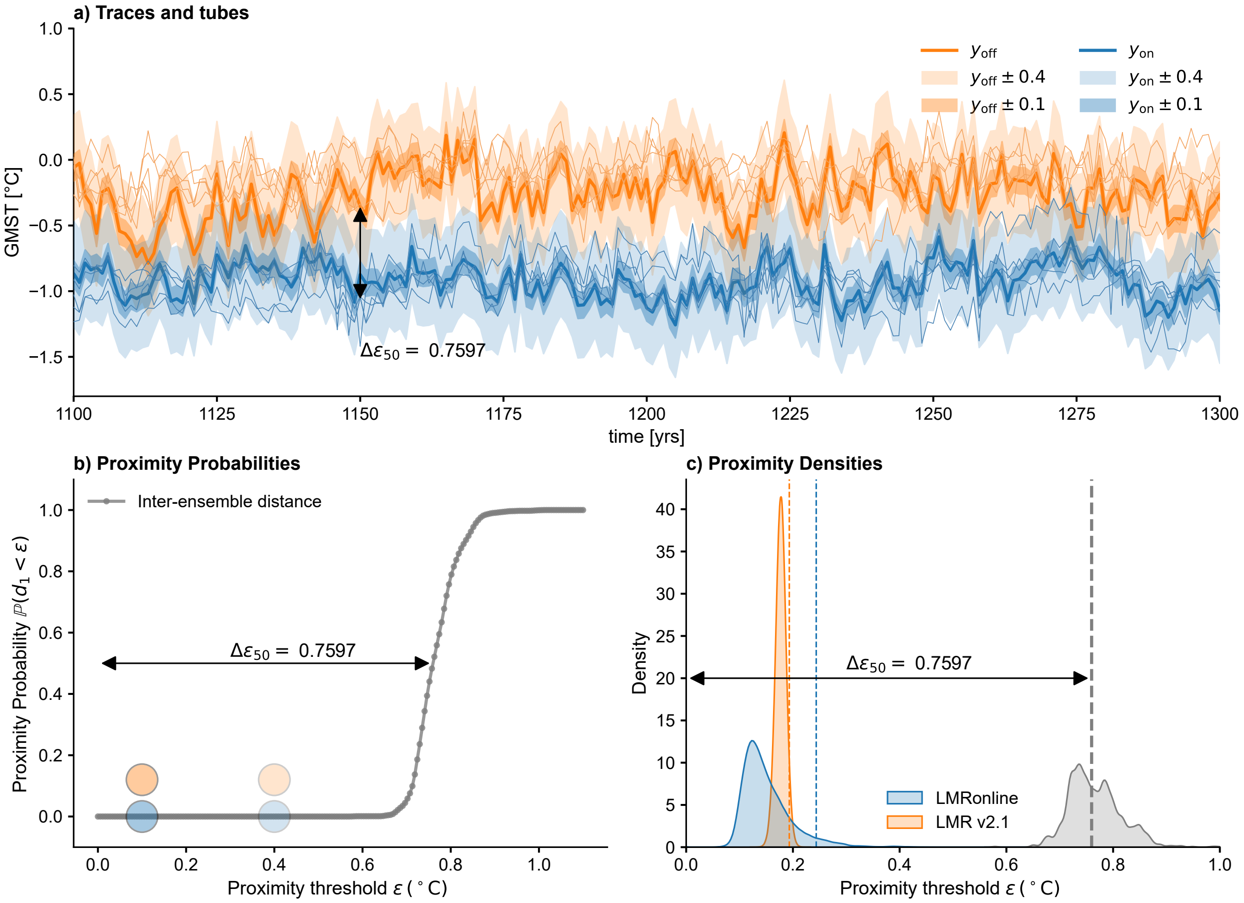

Fig. 9. **Plume distance schematic** for GMST in LMR v2.1 (orange) and LMRonline (blue), shifted downward by 0.75°C for illustrative purposes. a) 6 random traces from each ensemble (thin lines), two of which $(y_{\mathrm{on}}, y_{\mathrm{off}})$ are surrounded by "tubes" of size $\epsilon = \pm 0.1$ and $\epsilon = \pm 0.4$°C. By varying the width of this tube, one arrives at an estimate of proximity probabilities (b), whose median is the plume distance, $\Delta\epsilon_{50}$. Colored dots indicate the values of ϵ considered in a). The inter-ensemble distance (dark gray) can then be compared to intra-ensemble distances (panel c), for instance its 95% quantiles, indicated by colored, vertical dashed lines (one for each ensemble). The same plume distance $\Delta\epsilon_{50}$ is highlighted on all three panels.

than any of the LMRv2.1 resampled ensembles – as attested by its larger threshold (q_{95}) (Fig. 9, blue dashed line) and 2) the lowest plume distance across all ensembles occurs with CESM. Naturally, the results would vary somewhat depending on which quantile is chosen for the threshold. It is worth emphasizing that several measures could be taken to improve the comparison. In particular, Zhu et al. (2020) found that including only grid cells that correspond to the sites of the proxies used in LMRv2.1, and adjusting for seasonal biases, can substantially improve such a comparison.

	q_{95}	BCC	CCSM4	gl	s2	IPSL	MPI	CSIRO	HadCM3	CESM	GISS
LMRon	0.26	0.41	0.49	**0.23**	0.98	0.64	**0.35**	**0.29**	**0.22**	**0.15**	**0.29**
LMRoff, AR(2)	0.18	0.42	0.50	0.25	0.99	0.65	0.36	0.30	0.23	**0.16**	0.31
LMRoff, fGn	0.20	0.42	0.50	0.25	0.99	0.65	0.36	0.30	0.23	**0.16**	0.31
LMRoff, $f^{-\beta}$	0.19	0.42	0.50	0.25	0.99	0.65	0.36	0.30	0.23	**0.16**	0.31

TABLE 1. **Plume distance to PMIP3 past1000 simulations**. “BCC" stands for BCC_CSM1_1, “gl" for FGOALS_gl, “s2" for FGOALS_s2, and “IPSL" for IPSL_CM5A_LR, “MPI" for MPI_ESM_P and “GISS" for GISS-E2-R. (Dufresne et al. 2013; Giorgetta et al. 2013; Gordon et al. 2000; Otto-Bliesner et al. 2015; Rotstayn et al. 2012; Schmidt et al. 2012, 2006; Stevenson et al. 2019; Watanabe et al. 2011; Wu et al. 2014). q_{95} denotes the 95% quantile of each intra-ensemble proximity density. Numbers in bold indicate traces that are compatible with each ensemble (i.e. the 95% HDI of the ensemble-to-trace proximity density encompasses the q_{95} of the intra-ensemble distribution).

b. Quantifying similarity in the PAGES 2k (2019) ensemble

We now apply our framework to measure the consistency of the reconstructions from the Neukom et al. (2019) ensemble. This ensemble is composed of 7 GMST reconstructions using common inputs (Apr–Mar averages of a subset of proxies from the PAGES 2k Consortium (2017) compilation) and 7 different statistical methods, including a version of offline DA (Hakim et al. 2016). Each method provided a 1,000-member reconstruction ensemble to represent uncertainties. While Neukom et al. (2019) found great inter-method consistency at decacal to multi-decacal scales, centennial patterns were highly method-dependent, and it is worth asking how compatible they are with the “offline" DA product in this ensemble. The latter, despite using a similar methodology, is distinct from the LMRv2.1 solution (Fig. 1a), in that it uses a different selection of paleoclimate proxies, different proxy system models, and different settings for the offline DA algorithm. It is thus worth assessing its ensemble proximity to LMRv2.1.

Fig. 10 shows the individual ensembles (a–g), as well as the distribution of inter-ensemble distances from the offline DA solution (simply called “DA", as per the original paper’s terminology) in panel h. Their significance can be assessed by comparing to the intra-ensemble distance of the DA ensemble (denoted DA-DA), whose 95% quantile is marked by a vertical dashed line. Interestingly, the DA-DA and DA-LMRv2.1 distributions nearly coincide, and cluster around higher values than most other distributions. The only methods that show more than 5% of trajectories above the

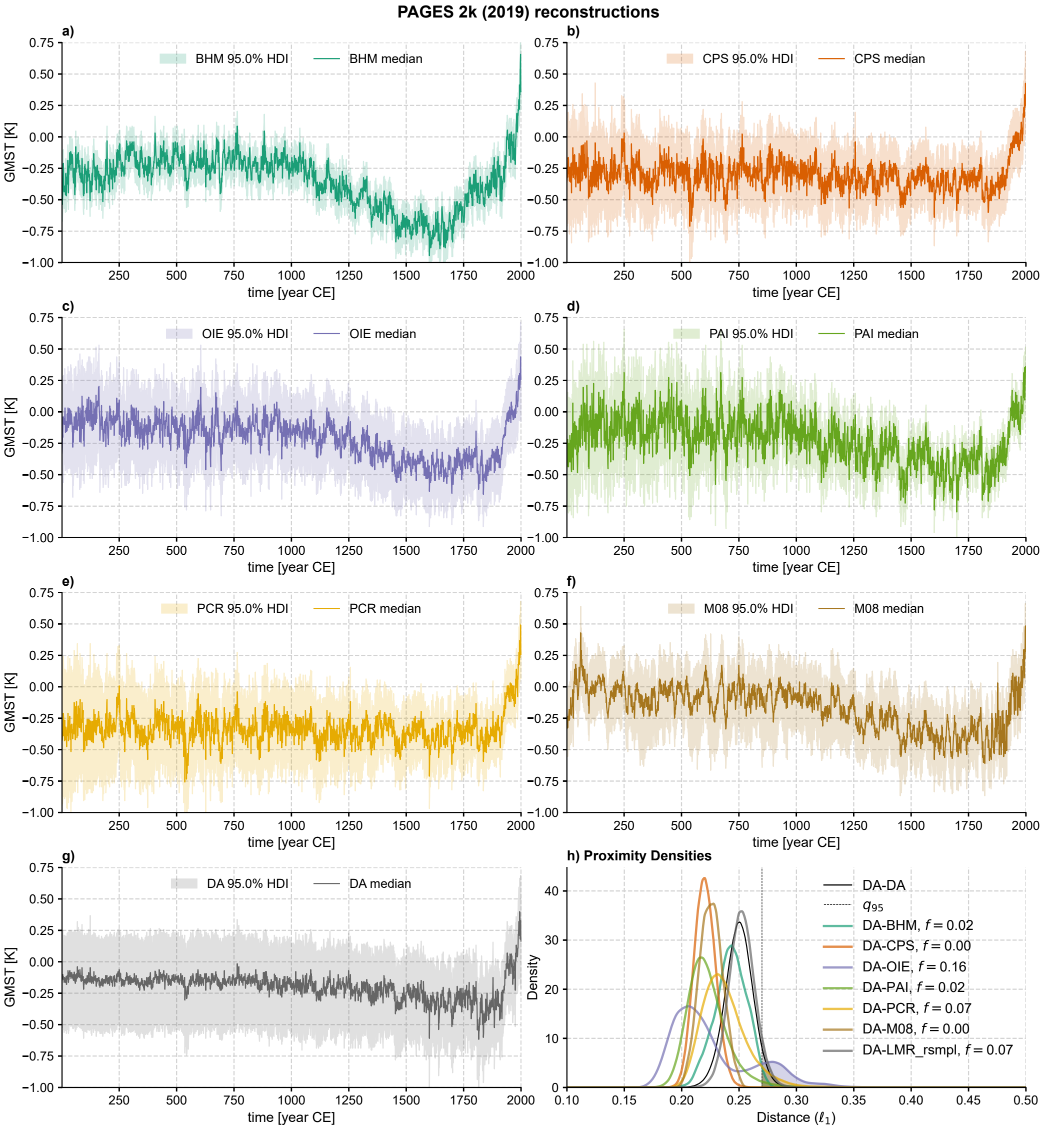

FIG. 10. **Comparisons with the Neukom et al. (2019) ensemble** Panels a-g correspond to reconstruction emsembles with the original methods: BHM, CPS, DA, OIE, PAI, PCR, and M08. Panel h shows the plume distance between the DA ensemble and the other 6 methods, as well as LMRv2.1. As before, the DA products are resampled according to a power law with $\beta = 0.92$ (see text for details). Distances are evaluated according to the ℓ^1 norm. The dashed line denotes the 95% quantile of the DA intra-ensemble distribution. f represents the fraction of each ensemble's trajectories that fall at a distance larger than this quantile. A number above 5% suggests incompatible ensembles.

95% quantile of the intra-ensemble (DA-DA) distance are LMRv2.1 (6%), PCR (7%), and OIE (16%). Only the latter may be said to be meaningfully different with a 95% threshold (also for a 99% threshold), unlike the other two. Thus, while there are important qualitative and quantitative differences among these 8 ensembles, this analysis only finds one method (OIE) to yield a meaningfully different estimate.

c. *Comparing reconstruction ensembles*

We now return to the example of Fig. 1c, showcasing the NHT reconstructions of Büntgen et al. (2021) (hereafter B21). To explore the impact of methodological choices in tree-ring reconstructions, B21 gathered 15 research groups to generate Northern Hemisphere summer temperature reconstructions from a common network of regional tree-ring width datasets. Despite the common inputs, their results vary notably in terms of spectral content and amplitude. How do they compare against those of another ensemble like LMRv2.1?

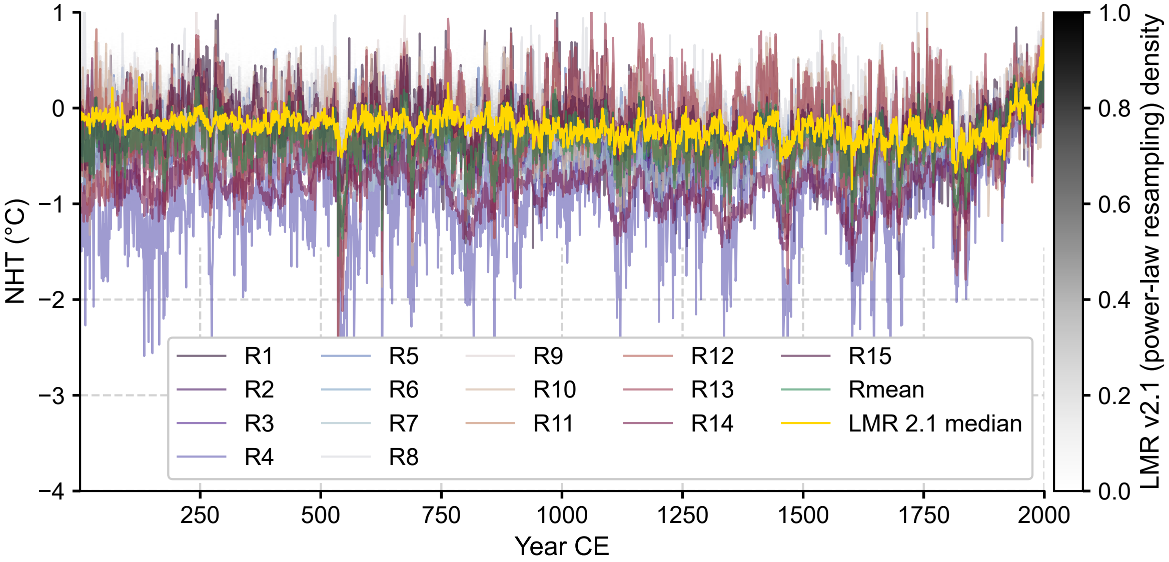

FIG. 11. Northern Hemisphere Surface Temperature (NHT) in the ensembles of Büntgen et al. (2021) (B21) and LMRv2.1 (Tardif et al. 2019). The LMR ensemble has been resampled according to a power law (Sect. b) to preserve scaling behavior. Its density (gray shading) is obscured by the large number of traces from B21.

The traces show (Fig. 11) that the LMR median displays very muted variability compared to most of the 15 ensemble members, or their ensemble mean ("Rmean"). However, the superposition of traces makes it difficult to judge which, if any, of the B21 reconstructions are compatible with the LMRv2.1 ensemble. While the aggregate measure that is the plume distance could help answer this question, one gets a more granular picture by plotting proximity densities themselves, as we did previously for the Neukom et al. (2019) ensemble. This comparison (Fig. 12) shows that 10 of the 15 B21 reconstructions are incompatible with the LMRv2.1 ensemble (their entire proximity densities lie beyond the 95% quantile of the LMRv2.1 intra-ensemble density); however, 5 of the reconstructions (R5, R6, R11, R12, R15) are compatible with LMRv2.1, and the majority of resampled LMRv2.1 traces are compatible with the B21 ensemble mean as well.

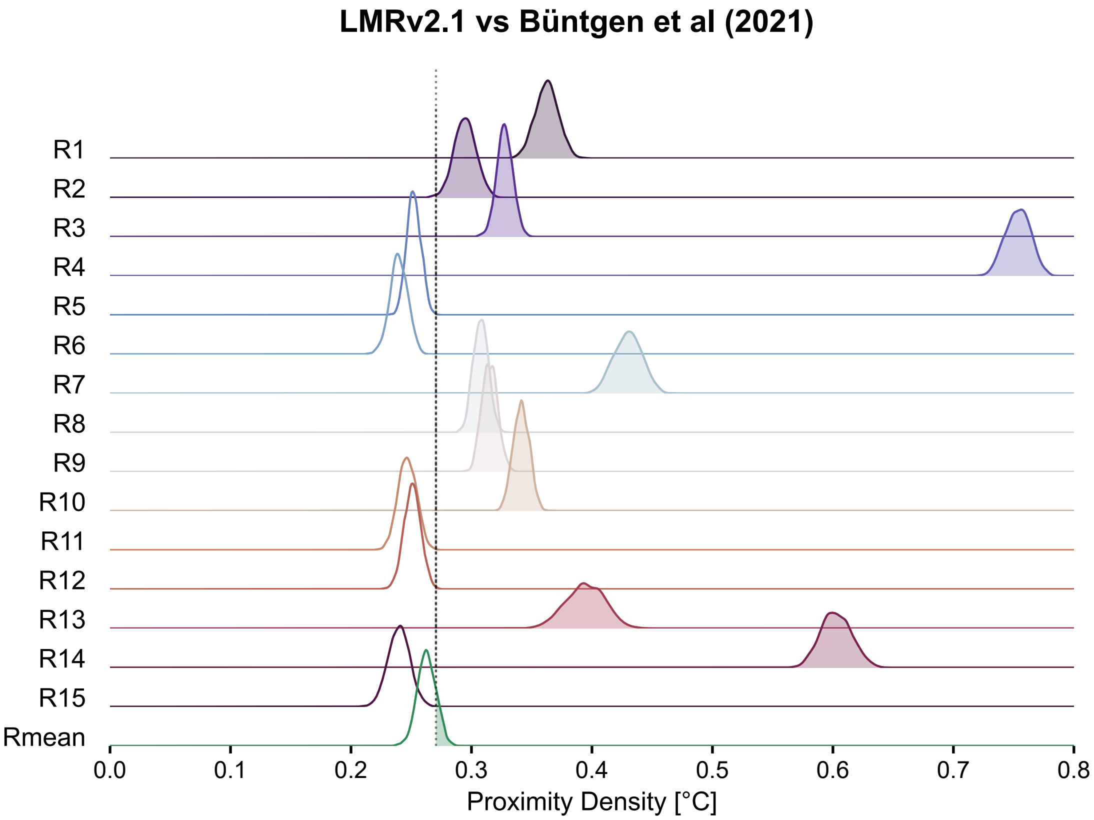

FIG. 12. **Proximity densities between LMRv2.1 and B21 ensembles**. Distances are evaluated according to the ℓ^1 norm, with respect to the LMR ensemble, and resampled according to a power law. The dashed line denotes the 95% quantile of the LMR intra-ensemble distribution. Shading denotes the fraction of each distribution that is incompatible with the baseline (LMRv2.1).

Note that this is necessarily a crude comparison; the 15 reconstructions of B21 used different target seasons, ranging from June–July to June–October, whereas LMR targets the annual mean, though its reliance on northern hemisphere tree-rings means that it is heavily biased towards a northern hemisphere summer season. However, the different target seasons imply that those two reconstruction ensembles have different target variances. In this context, it is notable that 5 of the B21 reconstructions are compatible with LMRv2.1.

6. Discussion

This article has addressed challenges with temporal diagnostics and comparisons using ensembles from data assimilation (DA). A key difficulty of this work was to devise a rigorous framework for comparing distributions of time-evolving trajectories to one another, or to a deterministic target (e.g. a model simulation). In researching this question, we were surprised to discover that there does not seem to exist a mathematical framework which would allow such comparisons in a meaningful way. The challenge for mathematicians who study long time series and continuous-time stochastic processes, and also for mathematically motivated time series scholars, is that when one fits a model for a single time series, probability theory dictates that the output of the fitting procedure must be specified as a probability measure on a space of time paths. Then, when the time series are long enough, two such models, even with slightly different specifications, may look irreconcilable in terms of what subsets of paths are accessible to each model from their respective measures, making a rigorous comparison all but impossible if one interprets a model as a probability measure.

This led to the formulation of a new metric called the plume distance, to measure distances between ensembles of traces, or between an ensemble and a trace. The notion of plume distance introduced here makes those comparisons very robust, and we argue, intuitive, as it takes on the properties of a norm, cast in the same units as the variable of interest (e.g. temperature). These tools were used to compare LMRv2.1 to LMRonline, to the PMIP3 `past1000` ensemble, the Neukom et al. (2019) ensemble, and to the Büntgen et al. (2021) ensemble.

In the case of the Last Millennium Reanalysis, as in all offline DA products, an essential problem is that the temporal behavior of ensemble perturbations from the mean is unconstrained by the method. We showed how the use of static priors can be partially overcome by adopting a parametric temporal model that leverages independent knowledge of the system, imparting a more realistic temporal behavior to the ensemble perturbations than naïve resampling, which whitens the ensemble time series. Coupled with the ensemble distance defined above, this resampling allowed for proper comparisons of traces and ensembles to an offline DA ensemble like LMRv2.1.

In that case, an online counterpart (Perkins and Hakim 2021) was available, and guided the choice of temporal model. The main advantage of online DA is that it propagates temporal information according to the dynamics of a physically-based model, providing constraints on the evolution of various climate variables, including those (like ocean heat content) that are only indirectly constrained by paleoclimate observations. In almost all scenarios imaginable, if an online DA estimate is available, it would be preferable to any offline DA estimate. However, in most cases involving offline DA, no such online counterpart is available. Indeed, for many deep-time applications, offline DA is the only practical option available at present. As such, we expect offline DA to endure for some time, and it is therefore critical to provide paleoclimatologists with useful strategies for diagnosing temporal properties of its output. The framework proposed here allows one to incorporate temporal characteristics of a climate variable (e.g. its power spectrum) and resample an offline DA ensemble in a way that allows for diagnostic climate applications.

One drawback of the present approach is that our construction inflates the temporal variance of solutions during periods of greater uncertainty, resulting in fluctuations that get larger at earlier times when proxy data are fewer. This feature is also apparent in the “damped" variance of the original offline DA ensemble members (e.g. Fig. 1a, colored curves), and is undesirable for some applications. Indeed, there is no *a priori* reason to assume that temporal variance of GMST over the Common Era is anything but constant, and many dendrochronological studies assume homoskedasticity (constant temporal variance) as part of the methodology (Cook 1990). A logical next question is how to construct solution traces that are consistent both with posterior uncertainties and homoskedastic internal climate variability. In analogy to “nested” reconstruction approaches like Composite-Plus-Scale (e.g. Bradley and Jones 1993), one approach could be to divide the reconstruction interval into a series of windows and, within each window, generate realizations of noise so that the spectrum of the total trace (i.e., ensemble mean plus noise) is equal to a target spectrum. Such an approach would account for the artificial heteroskedasticity (uneven variance over time) arising in the ensemble mean over time as a result of data availability. A challenge is that errors in the estimation of posterior errors or in the homoskedastic assumption could lead to situations where such traces cannot actually be found. Moreover, care must be taken to isolate the forced signal due to anthropogenic changes, which introduces heteroskedasticity of its own. Nevertheless, approaches that minimize variability artifacts arising from changing observing networks are necessary to test hypotheses about changing climate variability for all estimation techniques, including instrumental reanalyses.

Another important extension would be to generalize these ideas to spatial problems (e.g. comparing two climate fields from different reconstructions). Even without these constraints, this would require an adequate space-time model for climate fields, which is a frontier research problem. Doubly-sparse Gaussian processes (Axen et al. 2022) may provide relevant analytical results that could form the basis of useful resampling strategies.

In the meantime, what should users of offline DA products do? It is important to recall that the ensemble mean is robust, and in some cases sufficient to provide useful diagnostics (for instances, with composites such as those used to diagnose the response to volcanic events as in Zhu et al. (2022)). However, caution is essential with nonlinear diagnostics (e.g. variance), in which case resampling is essential. The code provided herein (`https://linked.earth/pens`) is appropriate for scalar variables, and can be applied for grid-point comparisons or spatial averages. A solution for spatio-temporal diagnostics of variance (e.g. empirical orthogonal functions) is an obvious point of focus for future work.

Finally, the distance framework introduced herein could be applied beyond paleoclimatology, in at least three areas:

1. Any ensemble-based forecast (or analysis) of environmental variables falls under this framework, so long as the focus is on a time-series (e.g. the NINO3.4 index for forecasts of El Niño-Southern Oscillation, or an air quality index over a metropolitan area). Although spatial variability is ostensibly of extreme importance, in practice many forecasts are issued as spatial averages over various scales, which present as plumes of time series, and are therefore amenable to this treatment.

2. In the field of stochastic finance, competing models for the time-evolution of prices of stocks and other financial instruments do not suffer from the difficulties described in Appendices B and C, but only in highly efficient and liquid markets (Hull 2017). In most other instances, e.g. emerging markets, our new distance framework could help explain statistically how market

participants make ad-hoc adjustments to implement financial risk management (Cartea et al. 2015; Yi et al. 2015), broadening its accessibility.

3. Nuclear physics models for the stability and radioactivity of heavy ions are complex mathematical questions, requiring severe numerical adjustments, often leading different research groups to making mutually inconsistent predictions. Recent solutions to these predicaments include model mixing strategies (Phillips et al. 2021), to the exclusion of any model comparisons, for lack of a systematic metric which could be viewed as fair. Drawing samples from different models for quantities of interest on the nuclear landscape (Neufcourt et al. 2019) would lead exactly into the framework of our distance tools, providing a systematic way of comparing models.

APPENDIX A

Timeseries models

Here we recall essential results of parametric time-series modeling, particularly the functional form of the spectral density and its dependence on model parameters.

a. $AR(p)$ models

A random process X is said to follow an autoregressive model of order p – that is, $AR(p)$ – if:

$$X_t - \mu = \sum_{k=1}^{p} \phi_k (X_{t-k} - \mu) + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2), \quad \phi_k \in \mathbb{R}. \tag{A1}$$

where $\mu = \mathbb{E}(X)$. Thus X_t depends only on the last p observations, plus an innovation term $\varepsilon_t \sim \mathcal{N}(0, {\sigma_\varepsilon}^2)$. The model's so called characteristic polynomial $\Phi(z) := z^p - \sum_{k=1}^{p} \phi_k z^{p-k}$ is useful to determine the behavior of X. If the equation $\Phi(z) = 0$ has all p of its (distinct complex) roots z_k strictly inside the complex unit circle, then the process X is stationary and its autocorrelation function $\rho(t)$ is a linear combination of the (complex) exponentials $(z_k)^t$. In this work the $AR(p)$ processes we consider are only the stationary ones.

The autocovariance function at lag $k > 0$ verifies the recurrence relation known as the Yule-Walker equations:

$$\gamma_t = \sum_{k=1}^{p} \varphi_k \gamma_{t-k} + \sigma_\varepsilon^2 \delta_{t,0} \tag{A2}$$

The solution is of the form:

$$\gamma_t = \sum_{k=1}^{p} \alpha_k z_k^t \tag{A3}$$

where the z_k's are the roots (assumed to be distinct) of the aforementioned characteristic polynomial equation $\Phi(z) = 0$, and $\alpha_1, \cdots, \alpha_p$ are arbitrary constants (Brockwell and Davis 2016), which can be determined by substituting (A3) into (A2) and solving this linear (Toeplitz), square system of equations. For the familiar stationary AR(1) model with $|\phi_1| < 1$, γ decays exponentially ($\gamma(t) = \phi_1^t$), which is emblematic of short-memory models. In practice, we use the `statsmodels` (Seabold and Perktold 2010) class `arima_process`[3] to fit this model and simulate from it.

b. Fractional Gaussian noise (fGn)

A paragon of long-memory models is the fractional Brownian motion (fBm), whose increments are the discrete-time fractional Gaussian noise (Qian 2003). A self-similar fractional Gaussian noise (fGn) process is a series of identically distributed Gaussian random variables $X_1, \cdots, X_n$ which are correlated over long ranges, in such a way that they are stable in distribution under fractional averaging:

$$\frac{X_1 + \cdots + X_N}{N^H} \sim X \tag{A4}$$

where $\sim$ means "distributed the same as" and $0 < H < 1$ is the Hurst exponent (Hurst 1951). The fGn's auto-covariance writes as:

$$\gamma(t) = \frac{1}{2}\left(|t+1|^{2H} + |t-1|^{2H} - 2|t|^{2H}\right) \tag{A5}$$

Such models are now ubiquitous in hydrology and other areas such as quantitative finance and internet traffic, and some have been argued to apply to climate behavior as well (Lovejoy and Schertzer 2013). For $H < 1$, such processes are stationary, though their memory decays much more slowly than autoregressive models (power law vs exponential). This slow decay exemplifies

[3] https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima_process.arma_generate_sample.html

long-range dependence (Beran 1994). In our work, we used the `FractionalGaussianNoise` class from the *stochastic* Python package[4] to generate such samples. In this work, H was calibrated from the scaling exponent β of the power spectrum of the online DA solution, using the relation $H = (\beta - 1)/2$.

c. Colored Noise

A third (and related) class of models centers on the spectrum itself. Many climate processes have been shown to exhibit a power law spectrum ($S(f) \propto f^{\beta}$) (Mitchell 1976; Pelletier 1998; Huybers and Curry 2006; Zhu et al. 2019; Franzke et al. 2020; Hébert et al. 2022), so it is natural to sample from such processes, which can be done through the `ColoredNoise` class of the *stochastic* Python package[5]. Such processes are related, but not identical to the fractional Gaussian noise described above.

Making use of the $t^{\alpha} \leftrightarrow f^{-(\alpha+1)}$ Fourier transform pair, one can express $\gamma(t) = t^{(\beta-1)}$. Therefore, the process is only stationary (with a decaying ACF) for $\beta < 1$, which corresponds to a Hurst exponent $H < 0.5$. Colored noise (power law) processes are therefore more general than fGn in the sense that they can represent longer-term memory, but are not necessarily stationary. For the range of parameters explored in this work, this distinction is immaterial, as all the processes investigated are stationary.

[4] https://stochastic.readthedocs.io/en/stable/noise.html#stochastic.processes.noise.FractionalGaussianNoise

[5] https://stochastic.readthedocs.io/en/stable/noise.html#colored-noise

APPENDIX B

Ensemble Likelihood: a failed attempt

Our work focused initially on establishing the likelihood of a trace $y(t)$ in the context of an ensemble $X(t)$, where X is sampled in M discrete traces. In the case of a Gaussian posterior ensemble – a reasonable approximation for the Last Millennium Reanalysis, for instance – the distribution of $X(t)$ for fixed t is fully characterized by its time-dependent mean $\mu(t)$ and standard deviation $\sigma(t)$, this likelihood is readily available at each time t:

$$\mathcal{L}_X(y,t) = \frac{1}{\sigma(t)}\varphi\left(\frac{y(t)-\mu(t)}{\sigma(t)}\right) \text{ where } \varphi(x) = \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}} \tag{B1}$$

While this poses no conceptual or analytical difficulty, the issue is numerical. Indeed, for a large temporal sample ($t = 0, \ldots 2000$), after accounting for serial correlations among the members of X, e.g. assuming that a good model to calibrate the empirical paths in X/σ is a stationary AR(p) model and then multiplying the vector $(y-\mu)/\sigma$ by the inverse of a square root of the fitted AR(p)'s auto-correlation matrix, the likelihood of an entire trace $\mathcal{L}_X(\mathbf{y})$ may be expressed as the product (where the auto-correlation matrix operation is suppressed for simplicity of notation, and the variances $\sigma^2(t)$ are assumed to be bounded below by some $\sigma_0^2 > 0$, as is the case with our data):

$$\mathcal{L}_X(\mathbf{y}) = \prod_{t=0}^{2000} \mathcal{L}_X(y,t) \leq \left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^{2001} \simeq 10^{-1839} \tag{B2}$$

where the order of magnitude above assumes that σ_0 is of order 1, which is also consistent with our data. The value on the right-hand side of (B2) is astronomically small (as a point of comparison, an upper bound on the number of atoms in the known universe is estimated to be around 10^{82}), and cannot be meaningfully distinguished from zero on any current machine architecture. As a result, any attempt to compute a likelihood ratio, even a log-likelihood, resulted in non-interpretable results. The issue here is that the size of the state vector is large enough to consider this question from the view point of continuous-time stochastic processes, but this requires making the same type of parametric assumptions made in Appendix A. Also see Appendix C for a discussion of where this non-interpretability most likely comes from. We prefer instead to work with the notion of plume distance, which is more intuitive, and preserves the units of the original variable (e.g. K for GMST).

APPENDIX C

Plume Distance

a. Definition

Here we flesh out the notion of plume distance described in Sect. 4. The idea is to give oneself a "tube" around a GMST simulation or similar trace, of size (e.g. radius) ϵ and shape determined by a norm on path space, such as the so-called ℓ^q norm, for some number $q \in [1,\infty]$. To fix ideas, for any $q \in [1,\infty)$, and a time span of T years, this tube around a trace $y = (y(t), t = 1, 2, \ldots, T)$, in the ℓ^q-norm, is the set of all trajectories x such that

$$\|y - x\|_q := \left(\sum_{t=1}^{T} |y(t) - x(t)|^q \right)^{1/q} \leq \epsilon. \tag{C1}$$

To get a sense of how compatible an ensemble is with a fixed trace, one may simply enumerate the number of ensemble trajectories that fit within ϵ of the target y, under the chosen norm. We do so via the following procedure. Given an ensemble X of trajectories, which is formed empirically of N paths x_i, we proceed as follows:

1. Compute the ℓ^q-norm distance $\|y - x\|_q$ between a trace y and each of the N members x_i in ensemble X.

2. Graph the distribution of distances $d_q(y, x_i) = \|y - x_i\|_q$, $i \in [1, \cdots, N]$, to choose a sensible range of ϵ parameters; this step, which can be performed by visual inspection of this distribution, is included to avoid considering values of ϵ which are extreme, saving computational effort.

3. Compute the proximity probability $\mathbb{P}(d_q(y, X) \leq \epsilon)$ as the proportion $n(\epsilon)/N$ of ensemble members that fit within the tube for a given set of ϵ parameters, where $n(\epsilon)$ is the number of members x_i of X which fit in that tube of size ϵ, i.e. the number of members x which satisfy the inequality condition $\|y - x\|_q \leq \epsilon$ in (C1).

4. Graph this proportion $n(\epsilon)/N$ as a function of ϵ.

In addition, we can use this proportion as a function of ϵ to compute simple, robust statistics of distance. This proportion is in fact a cumulative distribution function (CDF), meaning that it is a function which increases from 0 to 1 over the entire range of possible values ϵ. Consequently, for any number $p \in (0, 1)$, the value ϵ which leads to the value p for this CDF is immediately interpretable as the $100p$-th percentile. As shown in the main body of the article, any non-tail percentile of the difference between proximity probability curves, which is measured in °C of GMST, represents such a statistic of distance. A simple approach is to pick the absolute difference between the values of ϵ for which these CDFs intersect $p = 0.5$ (the median). It is interpreted as the most representative value, measured in the same units as y or X (°C in this article), for how far X is from the target y, and is thus the most natural benchmark upon which to base comparisons among ensembles.

To be clear, using the notation defined above, we define the *plume distance* $d(y, X)$ from the ensemble X to the trace y as the smallest value ϵ such that $\mathbb{P}(\|y - X\|_q \leq \epsilon) = n(\epsilon)/N$ equals or exceeds 0.5. Since this definition relies on the difference $y - X$, we may also say that the plume distance from X to y is equal to the plume distance from 0 to $X - y$, i.e. $d(0, X - y)$. This remark will be convenient below to explain the legitimacy of the distance properties of this plume distance.

Since the number N of members of the ensemble X is typically large, one typically finds that there exists a value ϵ such that $n(\epsilon)/N$ equals (almost) exactly ϵ (say, within an error less than $1/N$).

b. The plume distance as a norm

Let us show that this "plume distance" verifies the conditions of a usual distance. In fact, we will show more, that the plume distance, interpreted as the distance to the zero path of the difference $X - y$ between ensemble X and the trace y, is actually a norm for $X - y$, because in addition to the four usual axioms of a distance, it also preserves scaling by a positive constant. That is important because, while our plume distance is a measurement in °C, a change into a different unit of temperature should only scale the distance by the same unit conversion factor. We present the proof of these five properties in the next five bullet points, except that the proof of the 4th point, on the triangle inequality, is given after this list.

Zero : The distance from an object to itself is zero: if all the individual distances are 0, the distribution is a delta function centered at 0. More interesting is the case where thin subsets of the same ensemble are compared: we show empirical evidence in the supplement[6] that the distance between two subsamples of the same plume will be small, but finite, and that it tends to decrease as the ensemble size gets larger (i.e. as the full distribution is better sampled).

Positivity The distance between two distinct points is always positive, as the metric can only pick ϵ values that are positive-definite.

Symmetry The distance from X to Y is always the same as the distance from Y to X. This is guaranteed by taking the absolute value of the difference in proximity probabilities at any quantiles.

Triangle inequality To be a true distance, the triangle inequality needs to hold. Here, one must pause to realize that the triangle inequality should apply to the ensemble's difference with the fixed trace, i.e. $X - y$, not to the ensemble by itself. We already noted the plume distance $d(X, y)$ also equals $d(X - y, 0)$. Thus the triangle inequality we seek to prove is that, for two ensembles X, Y, then $d(X - y + Y - y, 0) \leq d(X - y, 0) + d(Y - y, 0)$. We provide a proof below. This requires deciding what it means to add two ensembles together; this is also elucidated below.

Scaling We must show that for any constant $c > 0$, $d(cX - cy, 0) = cd(X - y, 0)$. This is immediate because, if ϵ is the smallest value such that $\mathbb{P}(\|y - X\|_q \leq \epsilon)$ equals or exceeds 0.5, then $\mathbb{P}(\|cy - cX\|_q \leq c\epsilon)$ is the same probability as the previous one above, and thus it also equals or exceeds 0.5, and $c\epsilon$ is the smallest value on the right hand side in this probability that achieves this 0.5.

To prove the triangle inequality claimed above, let us assume that the plume distance for the two differences $X - y$ and $Y - y$ are attained exactly. Therefore let ϵ_1 and ϵ_2 be the two values such that

$$\mathbb{P}(\|y - X\|_q \leq \epsilon_1) = 0.5,$$

$$\mathbb{P}(\|y - Y\|_q \leq \epsilon_2) = 0.5.$$

[6] https://fzhu2e.github.io/pens/ug-examples.html

Thus by definition, ϵ_1 and ϵ_2 are the plume distances $d(X-y,0)$ and $d(Y-y,0)$. Also let ϵ be the value such that

$$\mathbb{P}(\|y-X+y-Y\|_q \leq \epsilon) = 0.5$$

so that by definition $d(X-y+Y-y,0) = \epsilon$. Next, as mentioned, we need a legitimate way to give a meaning to $y-X+y-Y$, the sum of the two ensemble deviations from y. To lighten the notation, we posit without loss of generality that $y = 0$. This means we must decide how to couple the two ensembles X,Y as probabilistic objects. Since each of X and Y is defined empirically as a set of equally likely trajectories, we only need to define a correspondence between trajectories of X and Y. The case where the number N of trajectories is an even number and is the same for X and Y is relatively straightforward, and we present the full proof in this case, leaving the general case for the interested reader, with the help of a comment at the end of this development.

Now, by definition of the plume distance, we know that there are exactly $N/2$ trajectories x in the ensemble X such that $\|x\|_q \leq \epsilon_1$. There are also exactly $N/2$ trajectories y in the ensemble Y such that $\|y\|_q \leq \epsilon_2$. The careful reader will excuse our slight abuse of nomenclature here, since now the letter y represents a generic member of the ensemble Y, whereas the trace target is understood as being equal to 0 without loss of generality. We couple the ensembles X and Y by assigning any fixed correspondence between each of those x's with the property $\|x\|_q \leq \epsilon_1$, to any one of the y's such that $\|y\|_q \leq \epsilon_2$. There are $(N/2)!$ ways of arranging this correspondence – any one of those ways is suitable. We repeat this procedure for setting a correspondence for the $N/2$ members x such that $\|x\|_q > \epsilon_1$ with those $N/2$ members y such that $\|y\|_q > \epsilon_2$.

With this correspondence (this coupling of the two ensembles) in place, the event $A := \left\{\|X\|_q \leq \epsilon_1\right\}$ is identical to the event $B := \left\{\|Y\|_q \leq \epsilon_2\right\}$. And these two identical events have probability equal to 0.5. Now for any $x \in A$, which corresponds to a specific $y \in B$, we have

$$\|x\|_q + \|y\|_q \leq \epsilon_1 + \epsilon_2.$$

However, since $\|\cdot\|_q$ is a norm, we have

$$\|x+y\|_q \leq \|x\|_q + \|y\|_q.$$

Combining these two, we get that for every $x \in A$ and its corresponding $y \in B$,

$$\|x+y\|_q \leq \epsilon_1 + \epsilon_2.$$

Therefore, on the common (empirical) probability space where X and Y are jointly defined, the number of members $x+y$ of the ensemble $X+Y$ such that the above inequality holds is at least equal to $N/2$, since that event contains A. Therefore,

$$\begin{aligned}\mathbb{P}(\|X+Y\|_q \leq \epsilon_1+\epsilon_2) \geq 0.5 \\ = \mathbb{P}(\|X+Y\|_q \leq \epsilon).\end{aligned}$$

Since CDFs are non-decreasing functions, this immediately implies that $\epsilon_1 + \epsilon_2 \geq \epsilon$, which by definition of the plume distance, means that

$$d(0, X+Y) \leq d(0,X) + d(0,Y)$$

This proves the triangle inequality, as announced, in the special case where the two ensembles have the same number of members N, by imposing a specific coupling among them. In the general case where the number N_1 of members of X may be, say, smaller than the number N_2 of members of Y, a coupling giving us the triangle inequality can also be devised. In this case, it is not possible to couple X and Y directly in such a way that $A = B$. The idea is first to identify the members of the event A as a subset of B, and then, for the members y' of B which are beyond the members of A, one must create an assignment of X which is consistent with norms being less than ϵ_1, but based on the fact that the corresponding y's have norms less than ϵ_2. The choice $X(y') = Y(y') \times \epsilon_1/\epsilon_2$ works, and leads to a situation that brings us back to the case where $N_1 = N_2 = N$ which was treated above. The details are left to the interested reader.

c. Robustness

Having established the triangle inequality for the norm on ensemble space which is the plume distance $d(0, X-y)$ defined as the 50th percentile of the proximity probability from X to y, we can return to the discussion of how robust this definition is. We have noted in the main body of the paper that the differences of these percentiles, for two traces compared to a benchmark ensemble, are not only robust across benchmark models of the offline LMR, but are also robust across all tube shapes, even though ℓ^q tubes for high-dimensional models are known mathematically to have drastically differing shapes. This may be surprising to those well aware of the non-equivalence of norms in infinite-dimensional linear spaces. However, it reflects a deep result in probability theory which was established in the last decades for Gaussian stochastic processes. We explain this here briefly, to shed light on the broader question of how to compare a trace and a model or ensemble of trajectories.

In our attempt to produce a likelihood-based notion of proximity or consistency of a single trajectory to a model, we investigated the appropriateness of the so-called small-ball probability (SmBP) in the theory of stochastic processes. The basic version of SmBP is the following. Consider a stochastic process X indexed by time, with mean equal to 0, such as an $AR(p)$ process, or a continuous-time process, e.g. the Ornstein-Uhlenbeck (OU), which is the high-frequency limiting process of AR(1). Let $\epsilon > 0$ be a given radius. The basic SmBP of X is the limiting behavior of the probability that X remains within the distance ϵ from the constant path at 0. This probability is $\mathbb{P}(||X|| \leq \epsilon)$, where the norm is up to the user to choose, for instance an ℓ^q norm. For Gauss-Markov processes, including OU and AR(1), it typically behaves like $\exp(-c/\epsilon^2)$ where c is a constant that depends on the type of process and on the norm used, while for other processes the behavior varies. For fractional Brownian motion with Hurst parameter H, for instance, the ϵ^2 is replaced by $\epsilon^{1/H}$ (Li and Shao 2001). The SmBP around a trace y which is different from 0 turns out to be a non-trivial question in many cases (Bongiorno and Goia 2017). However, for mean-zero Gaussian processes, the SmBP around a non-zero trace y behaves asymptotically like the same SmBP around 0, times a term $L(y)$ which does not depend on ϵ, and depends instead on the so-called large deviations behavior of X, in the sense that $L(y)$ is determined by the norm of y in the so-called reproducing kernel Hilbert space (RKHS) of X, regardless of what norm is used to defined the SmBP. Details of this result are in Section 3.1 of Li and Shao (2001).

This extraordinary property of Gaussian processes shows that the intuitive notion of how likely it is for a model to be within a "distance" ϵ of a trace, can be decomposed as the product of the SmBP around 0, interpreted as a volume element with a prescribed behavior for small ϵ which is not connected to the nature of the trace y, times a likelihood $L(y)$ of the trace which is the same no matter what notion of distance is chosen, and does not depend on ϵ. This theory points to SmBP and the corresponding likelihood as appropriate ways of comparing fixed paths with models. While we were not able to show in practice that this notion of likelihood is a robust statistic for our models, ensembles, and traces, the fact that the SmBP likelihood does not depend on the type of norm or distance being considered, is confirmed in our analysis of the plume distance, which is precisely the macroscopic version of SmBP, when ϵ is not sent to 0. The proposed plume distance statistic ($\Delta\epsilon$) is quite insensitive to the choice of the norm ℓ^q, as predicted asymptotically as $\epsilon \to 0$ by Theorem 3.1 in Li and Shao (2001).

We also noted that the plume distance $\Delta\epsilon_{50}$ is insensitive to the type of model being used, whether an $AR(p)$, or a power-law ACF, or an fGn (Appendix A), or the empirical non-parametric model defined by the ensemble itself. This is indicative of the idea that the distinctions between the various models' RKHS's are not prominent at the non-asymptotic scale defined by our statistic $\Delta\epsilon_{50}$. The consistency between a trace and a model appears to be driven by non-parametric properties of the trace as it compares to a reasonable cloud of trajectories. This phenomenon is one of the behavior of stochastic processes at a mesoscopic scale. It is not covered in the theoretical literature on stochastic processes because that area of research focuses more on asymptotics, or on global properties. It is worthy of further investigation in practice and in theory.

d. Necessity

We finish with a brief technical note on the necessity of introducing this new notion of plume distance. That is, we discuss the inappropriateness of other ways to measure the consistency or proximity between models and/or traces. We focus on the popular tool of Kullback-Leibler (K-L) divergence (see for instance Bishop (2006)), though some of these elements apply to other common metrics such as Continuous-Ranked Probability Scores (CRPS, Matheson and Winkler (1976); Gneiting and Katzfuss (2014)). The K-L divergence $D_{\mathrm{KL}}(P|Q)$ from a benchmark model Q to an alternative proposal P for a model is computed as the entropy of the alternative model relative to the benchmark. This quantity represents an information content, and is not a norm in the physical space of GMSTs. Moreover, it requires the benchmark to be a model rather than a single trace. These two features make it less appropriate than a norm in physical space like the plume distance, which can draw comparisons to a single trajectory.

There is yet a more serious drawback to K-L divergence. The relative entropy between two models can only be computed if the so-called Radon-Nikodym derivative of the proposal model with respect to the benchmark model can be computed unambiguously. This derivative only exists unambiguously if the proposal model has the property of being absolutely continuous with respect to the benchmark. This means that an event has a zero chance of occurring for the proposal model as soon as its chance is zero for the benchmark. In the limit of large number of observations, our time series models of interest, like $AR(p)$, are known to converge to continuous-time stochastic processes. For instance, as mentioned, the $AR(1)$ time series converges to an OU process, which is the solution of a linear stochastic differential equation driven by a Brownian motion. The problem is that, far from having two OU models, for instance, be absolutely continuous with respect to each other, unless the models are identical or have identical driving uncertainty intensity (which would never happen in practice for models or ensembles coming from different research teams), they are at the very other extreme: they are singular with respect to each other, i.e. the trajectories that support one of the models have no chance of occurring under the other model (to be precise, the smallest closed set of trajectories that supports one model has zero probability of occurrence under the other model). This implies that the Radon-Nikodym derivative of one OU with respect to another OU does not exist (unless they share the exact same noise intensity), thus the K-L divergence from one OU to another is not well defined.

Some authors propose an artificial measure-theoretic fix to this conundrum by suggesting that one take the Radon-Nikodym derivatives of either of the two models P, Q with respect to the mixture model M where each one of P and Q has a 50% chance of occurring, namely $M := (P+Q)/2$, and using those derivatives in the definition of the K-L divergence. In the explanation that follows, we will often use the term "density" when speaking of Radon-Nikodym derivatives, when this is unambiguous. The idea to use M stems from the original work of Kullback and Leibler (Kullback and Leibler 1951), where a symmetrization of their divergence is proposed, leading to the idea of symmetrizing the reference measure. That idea produces the so-called Jensen-Shannon divergence, formally $D_{\mathrm{JS}}(P,Q) := D_{\mathrm{KL}}(P|M) + D_{\mathrm{KL}}(Q|M)$ which coincides locally (up to a universal proportionality factor) with the Fisher information metric, resulting in a symmetric statistic (Nielsen 2019). The same idea leads to defining $D_{\mathrm{KL}}(P|Q)$ by expressing the entropy of P with respect to Q by simply using the densities of both P and Q relative to M (Bishop 2006). Those densities exist, but when the measures P, Q are singular with respect to each other, the densities are supported on disjoint portions of the space where M is defined, leading to an undefined $D_{\mathrm{KL}}(P|Q)$. Each corresponding Radon-Nikodym derivative would be non-zero exactly when the other is equal to 0, leading to an expression of the form $-\infty+\infty$, i.e. $\ln(0/0)$, which is undefined. Therefore the theoretical fix of relying on densities with respect to M does not apply to mutually singular models, as one gets for two OU processes with different noise intensities, or more broadly for any pair of long-horizon limits of AR(p) models with even minor differences in auto-regressive coefficients. We believe that this phenomenon leads to K-L divergences for two different time series models which are extremely unstable as the number of time steps climbs into the hundreds and thousands, and can be arbitrarily large in absolute value, leading to a meaningless metric. We think this is precisely the same phenomenon which we observed numerically with our own data, and which we described in Appendix B.

The same phenomenon of an undefined K-L divergence will occur when using densities relative to any mixture of P and Q, not merely the 50/50 mixture M, anytime P and Q are mutually singular. It is important to note that when the time series under consideration are of moderate length (dozens of time steps rather than hundreds or thousands), the use of M, or of other mixtures, as a benchmark, would typically not suffer from the issues described above, since any two legitimate models P, Q describing the same time series data would not be close to mutually singular, and thus the densities of P and Q with respect to $M = (P+Q)/2$ would share a common support of sufficient girth, so to speak, to allow a meaningful comparison from Q to P. As reported in Appendix B, it does not appear that our data allows us to be close to such a scenario.

Acknowledgments. We thank 3 anonymous reviewers for constructive comments on the submitted draft of this manuscript. JEG and FZ acknowledge funding from NSF grants 1948822 and 2126510. FV acknowledges funding from NSF grant DMS 2311306. GJH acknowledges funding from NSF grant 2202526 and Heising-Simons Foundation grant 2023-4715. This material is based upon work supported by the NSF National Center for Atmospheric Research, which is a major facility sponsored by the U.S. National Science Foundation under Cooperative Agreement No. 1852977.

Availability Statement

The Python code to reproduce the key figures of this work is available at `https://linked.earth/pens`.

References

Acevedo, W., B. Fallah, S. Reich, and U. Cubasch, 2017: Assimilation of pseudo-tree-ring-width observations into an atmospheric general circulation model. *Climate of the Past*, **13 (5)**, 545–557, https://doi.org/10.5194/cp-13-545-2017.

Amrhein, D. E., C. Wunsch, O. Marchal, and G. Forget, 2018: A global glacial ocean state estimate constrained by upper-ocean temperature proxies. *Journal of Climate*, **31 (19)**, 8059–8079, https://doi.org/10.1175/JCLI-D-17-0769.1.

Annan, J. D., and J. C. Hargreaves, 2012: Identification of climatic state with limited proxy data. *Clim. Past*, **8 (4)**, 1141–1151, https://doi.org/10.5194/cp-8-1141-2012.

Axen, S. D., A. Gessner, C. Sommer, N. Weitzel, and A. Tejero-Cantero, 2022: Spatiotemporal modeling of european paleoclimate using doubly sparse gaussian processes. arXiv, URL https://arxiv.org/abs/2211.08160, https://doi.org/10.48550/ARXIV.2211.08160.

Beran, J. ., 1994: *Statistics for Long-Memory Processes*. Chapman & Hall.

Bishop, C. M., 2006: *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Bongiorno, E. G., and A. Goia, 2017: Some insights about the small ball probability factorization for hilbert random elements. *Statistica Sinica*, 1949–1965, https://doi.org/10.5705/ss.202016.0128.

Braconnot, P., S. P. Harrison, M. Kageyama, P. J. Bartlein, V. Masson-Delmotte, A. Abe-Ouchi, B. Otto-Bliesner, and Y. Zhao, 2012: Evaluation of climate models using palaeoclimatic data. *Nature Clim. Change*, **2 (6)**, 417–424, https://doi.org/10.1038/nclimate1456.

Bradley, R. S., and P. D. Jones, 1993: ’Little Ice Age’ summer temperature variations: their nature and relevance to recent global warming trends. *The Holocene*, **3 (4)**, 367–376.

Brockwell, P. J., and R. A. Davis, 2016: *Introduction to Time Series and Forecasting*. Springer Texts in Statistics, Springer International Publishing, https://doi.org/10.1007/978-3-319-29854-2, URL https://doi.org/10.1007/978-3-319-29854-2.

Büntgen, U., and Coauthors, 2020: Prominent role of volcanism in common era climate variability and human history. *Dendrochronologia*, **64**, 125 757, https://doi.org/10.1016/j.dendro.2020.125757.

Büntgen, U., and Coauthors, 2021: The influence of decision-making in tree ring-based climate reconstructions. *Nature Communications*, **12 (1)**, 3411, https://doi.org/10.1038/s41467-021-23627-6.

Carrassi, A., M. Bocquet, L. Bertino, and G. Evensen, 2018: Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, **9 (5)**, e535, https://doi.org/10.1002/wcc.535.

Cartea, Á., S. Jaimungal, and J. Penalva, 2015: *Algorithmic and high-frequency trading*. Cambridge University Press.

Cook, L. A., E. R.and Kairiukstis, Ed., 1990: *Methods of dendrochronology: applications in the environmental sciences*. 394 pp, https://doi.org/ISBNISBN0-7923-0586-8.

Dee, S. G., and N. J. Steiger, 2022: ENSO's Response to Volcanism in a Data Assimilation-Based Paleoclimate Reconstruction Over the Common Era. *Paleoceanography and Paleoclimatology*, **37 (3)**, e2021PA004 290, https://doi.org/10.1029/2021PA004290.

Dufresne, J.-L., and Coauthors, 2013: Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. *Climate Dynamics*, **40 (9-10)**, 2123–2165, https://doi.org/10.1007/s00382-012-1636-1.

Erb, M. P., J. Emile-Geay, G. J. Hakim, N. Steiger, and E. J. Steig, 2020: Atmospheric dynamics drive most interannual U.S. droughts over the last millennium. *Science Advances*, **6 (32)**, eaay7268, https://doi.org/10.1126/sciadv.aay7268.

Erb, M. P., N. P. McKay, N. Steiger, S. Dee, C. Hancock, R. F. Ivanovic, L. J. Gregoire, and P. Valdes, 2022: Reconstructing holocene temperatures in time and space using paleoclimate data assimilation. *Climate of the Past*, **18 (12)**, 2599–2629, https://doi.org/10.5194/cp-18-2599-2022.

Fraedrich, K., U. Luksch, and R. Blender, 2004: $1/f$ model for long-time memory of the ocean surface temperature. *Phys. Rev. E*, **70**, 037 301, https://doi.org/10.1103/PhysRevE.70.037301.

Franke, J., S. Brönnimann, J. Bhend, and Y. Brugnara, 2017: A monthly global paleo-reanalysis of the atmosphere from 1600 to 2005 for studying past climatic variations. *Scientific Data*, **4**, 170 076 EP –.

Franzke, C. L. E., and Coauthors, 2020: The structure of climate variability across scales. *Reviews of Geophysics*, **58 (2)**, https://doi.org/10.1029/2019rg000657.

Fredriksen, H.-B., and K. Rypdal, 2016: Spectral characteristics of instrumental and climate model surface temperatures. *Journal of Climate*, **29 (4)**, 1253–1268, https://doi.org/10.1175/JCLI-D-15-0457.1.

Gebhardt, C., N. Kuehl, A. Hense, and T. Litt, 2008: Reconstruction of Quaternary temperature fields by dynamically consistent smoothing. *Climate Dynamics*, **30 (4)**, 421–437.

Giorgetta, M. A., and Coauthors, 2013: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *Journal of Advances in Modeling Earth Systems*, **5 (3)**, 572–597, https://doi.org/10.1002/jame.20038.

Gneiting, T., and M. Katzfuss, 2014: Probabilistic forecasting. *Annual Review of Statistics and Its Application*, **1 (1)**, 125–151, https://doi.org/10.1146/annurev-statistics-062713-085831.

Goosse, H., E. Crespin, A. de Montety, M. E. Mann, H. Renssen, and A. Timmermann, 2010: Reconstructing surface temperature changes over the past 600 years using climate model simulations with data assimilation. *Journal of Geophysical Research: Atmospheres*, **115 (D9)**, https://doi.org/10.1029/2009JD012737.

Goosse, H., H. Renssen, A. Timmermann, R. S. Bradley, and M. E. Mann, 2006: Using paleoclimate proxy-data to select optimal realisations in an ensemble of simulations of the climate of the past millennium. *Climate Dynamics*, **27 (2-3)**, 165–184.

Gordon, C., C. Cooper, C. A. Senior, H. Banks, J. M. Gregory, T. C. Johns, J. F. B. Mitchell, and R. A. Wood, 2000: The simulation of sst, sea ice extents and ocean heat transports in a version of the hadley centre coupled model without flux adjustments. *Clim. Dyn.*, **16 (2/3)**, 147–168.

Hakim, G. J., J. Emile-Geay, E. J. Steig, D. Noone, D. M. Anderson, R. Tardif, N. Steiger, and W. A. Perkins, 2016: The last millennium climate reanalysis project: Framework and first results. *Journal of Geophysical Research: Atmospheres*, **121 (12)**, 2016JD024 751, https://doi.org/10.1002/2016JD024751.

Hasselman, K., 1976: Stochastic climate models. part i. theory. *Tellus*, **28**, 473–485, https://doi.org/10.3402/tellusa.v28i6.11316.

Hébert, R., U. Herzschuh, and T. Laepple, 2022: Millennial-scale climate variability over land overprinted by ocean temperature fluctuations. *Nature Geoscience*, **15 (11)**, 899–905, https://doi.org/10.1038/s41561-022-01056-4.

Hull, J. C., 2017: *Options, futures, and other derivatives*. 10th ed., Pearson.

Hurst, H. E., 1951: Long term storage capacities of reservoirs. *Trans. ASCE*, **116**, 776–808.

Huybers, P., and W. Curry, 2006: Links between annual, milankovitch and continuum temperature variability. *Nature*, **441 (7091)**, 329–332.

Hyndman, R. J., 1996: Computing and graphing highest density regions. *The American Statistician*, **50 (2)**, 120–126, https://doi.org/10.1080/00031305.1996.10474359.

IPCC, 2007: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

IPCC, 2021: Summary for policymakers. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, Eds., Cambridge University Press.

Jones, J., and M. Widmann, 2004: Reconstructing Large-scale Variability from Palaeoclimatic Evidence by Means of Data Assimilation Through Upscaling and Nudging (DATUN). *The KIHZ project: Towards a Synthesis of Holocene Proxy Data and Climate Models*, H. Fischer,

T. Kumke, G. Lohmann, G. Flösser, H. Miller, H. von Storch, and J. Negendank, Eds., Springer, Heidelberg, Berlin, New York, 171–193.

Judd, K., C. A. Reynolds, T. E. Rosmond, and L. A. Smith, 2008: The geometry of model error. *Journal of the Atmospheric Sciences*, **65 (6)**, 1749 – 1772, https://doi.org/10.1175/2007JAS2327.1.

Judd, K., and L. A. Smith, 2004: Indistinguishable states ii: The imperfect model scenario. *Physica D: Nonlinear Phenomena*, **196 (3)**, 224–242, https://doi.org/https://doi.org/10.1016/j.physd.2004.03.020.

Khider, D., J. Emile-Geay, F. Zhu, A. James, J. Landers, V. Ratnakar, and Y. Gil, 2022: Pyleoclim: Paleoclimate Timeseries Analysis and Visualization With Python. *Paleoceanography and Paleoclimatology*, **37 (10)**, e2022PA004 509, https://doi.org/10.1029/2022PA004509.

King, J. M., K. J. Anchukaitis, J. E. Tierney, G. J. Hakim, J. Emile-Geay, F. Zhu, and R. Wilson, 2021: A data assimilation approach to last millennium temperature field reconstruction using a limited high-sensitivity proxy network. *Journal of Climate*, 1–64, https://doi.org/10.1175/JCLI-D-20-0661.1.

Kirchner, J. W., 2005: Aliasing in $1/f^\alpha$ noise spectra: Origins, consequences, and remedies. *Physical Review E*, **71 (6)**, 066 110–, https://doi.org/10.1103/PhysRevE.71.066110.

Kullback, S., and R. A. Leibler, 1951: On information and sufficiency. *The Annals of Mathematical Statistics*, **22 (1)**, 79–86.

Laepple, T., and P. Huybers, 2014: Global and regional variability in marine surface temperatures. *Geophysical Research Letters*, **41 (7)**, 2528–2534, https://doi.org/10.1002/2014GL059345.

Laepple, T., and Coauthors, 2023: Regional but not global temperature variability underestimated by climate models at supradecadal timescales. *Nature Geoscience*, **16 (11)**, 958–966, https://doi.org/10.1038/s41561-023-01299-9.

Li, W. V., and Q. M. Shao, 2001: *Gaussian processes: Inequalities, small ball probabilities and applications*, Vol. 19, 533–597. Elsevier, https://doi.org/10.1016/S0169-7161(01)19019-X, URL https://www.sciencedirect.com/science/article/pii/S016971610119019X.

Lovejoy, S., 2015: A voyage through scales, a missing quadrillion and why the climate is not what you expect. *Climate Dynamics*, **44 (11-12)**, 3187–3210, https://doi.org/10.1007/s00382-014-2324-0.

Lovejoy, S., and D. Schertzer, 2013: *The Weather and Climate: Emergent Laws and Multifractal Cascades*. Cambridge University Press, URL https://books.google.com/books?id=SeBjgLD43IIC.

Maraun, D., H. W. Rust, and J. Timmer, 2004: Tempting long-memory - on the interpretation of dfa results. *Nonlinear Processes in Geophysics*, **11 (4)**, 495–503, https://doi.org/10.5194/npg-11-495-2004.

Masson-Delmotte, V., and Coauthors, 2013: Information from Paleoclimate Archives. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley, Eds., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 383–464, https://doi.org/10.1017/CBO9781107415324.013.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Management science*, **22 (10)**, 1087–1096.

Mitchell, J. M., 1976: An overview of climatic variability and its causal mechanisms. *Quaternary Research*, **6 (4)**, 481–493, https://doi.org/10.1016/0033-5894(76)90021-1.

Neufcourt, L., Y. Cao, W. Nazarewicz, E. Olsen, and F. Viens, 2019: Neutron Drip Line in the Ca Region from Bayesian Model Averaging. *Phys. Rev. Lett.*, **122**, 062 502, https://doi.org/10.1103/PhysRevLett.122.062502.

Neukom, R., N. Steiger, D. Kaufman, and M. Grosjean, 2022: Inconsistent comparison of temperature reconstructions over the Common Era. *Dendrochronologia*, **74**, 125 965, https://doi.org/10.1016/j.dendro.2022.125965.

Neukom, R., and Coauthors, 2019: Consistent multidecadal variability in global temperature reconstructions and simulations over the Common Era. *Nature Geoscience*, **12 (8)**, 643–649, https://doi.org/10.1038/s41561-019-0400-0.

Nielsen, F., 2019: On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means. *Entropy*, **21 (5)**, https://doi.org/10.3390/e21050485.

Osman, M. B., J. E. Tierney, J. Zhu, R. Tardif, G. J. Hakim, J. King, and C. J. Poulsen, 2021: Globally resolved surface temperatures since the Last Glacial Maximum. *Nature*, **599 (7884)**, 239–244, https://doi.org/10.1038/s41586-021-03984-4.

Otto-Bliesner, B. L., and Coauthors, 2015: Climate variability and change since 850 CE: An ensemble approach with the community earth system model. *Bull. Amer. Meteor. Soc.*, **97 (5)**, 735–754, https://doi.org/10.1175/BAMS-D-14-00233.1.

PAGES 2k Consortium, 2017: A global multiproxy database for temperature reconstructions of the Common Era. *Scientific Data*, **4**, 170 088 EP, https://doi.org/10.1038/sdata.2017.88.

Pelletier, J. D., 1998: The power spectral density of atmospheric temperature from time scales of 10^{-2} to 10^6 yr. *Earth and Planetary Science Letters*, **158 (3)**, 157–164, https://doi.org/10.1016/S0012-821X(98)00051-X.

Perkins, W. A., and G. Hakim, 2020: Linear inverse modeling for coupled atmosphere-ocean ensemble climate prediction. *Journal of Advances in Modeling Earth Systems*, **12 (1)**, e2019MS001 778, https://doi.org/10.1029/2019MS001778.

Perkins, W. A., and G. J. Hakim, 2017: Reconstructing paleoclimate fields using online data assimilation with a linear inverse model. *Climate of the Past*, **13 (5)**, 421–436, https://doi.org/10.5194/cp-13-421-2017.

Perkins, W. A., and G. J. Hakim, 2021: Coupled atmosphere–ocean reconstruction of the last millennium using online data assimilation. *Paleoceanography and Paleoclimatology*, **36 (5)**, e2020PA003 959, https://doi.org/10.1029/2020PA003959.

Phillips, D. R., and Coauthors, 2021: Get on the BAND Wagon: a Bayesian framework for quantifying model uncertainties in nuclear dynamics. *Journal of Physics G: Nuclear and Particle Physics*, **48 (7)**, 072 001, https://doi.org/10.1088/1361-6471/abf1df.

Pope, V. D., M. L. Gallani, P. R. Rowntree, and R. A. Stratton, 2000: The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3. *Clim. Dyn.*, **16**, 123–146.

Qian, H., 2003: *Fractional Brownian Motion and Fractional Gaussian Noise*, 22–33. Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/3-540-44832-2{_}2.

Rotstayn, L. D., S. J. Jeffrey, M. A. Collier, S. M. Dravitzki, A. C. Hirst, J. I. Syktus, and K. K. Wong, 2012: Aerosol- and greenhouse gas-induced changes in summer rainfall and circulation in the Australasian region: a study using single-forcing climate simulations. *Atmos. Chem. Phys.*, **12 (14)**, 6377–6404, https://doi.org/10.5194/acp-12-6377-2012.

Schmidt, G. A., and Coauthors, 2006: Present-Day Atmospheric Simulations Using GISS ModelE: Comparison to In Situ, Satellite, and Reanalysis Data. *Journal of Climate*, **19 (2)**, 153–192, https://doi.org/10.1175/JCLI3612.1.

Schmidt, G. A., and Coauthors, 2012: Climate forcing reconstructions for use in pmip simulations of the last millennium (v1.1). *Geoscientific Model Development*, **5 (1)**, 185–191, https://doi.org/10.5194/gmd-5-185-2012.

Seabold, S., and J. Perktold, 2010: statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Shoji, S., A. Okazaki, and K. Yoshimura, 2022: Impact of proxies and prior estimates on data assimilation using isotope ratios for the climate reconstruction of the last millennium. *Earth and Space Science*, **9 (5)**, e2020EA001 618, https://doi.org/10.1029/2020EA001618.

Singh, H. K. A., G. J. Hakim, R. Tardif, J. Emile-Geay, and D. C. Noone, 2018: Insights into Atlantic multidecadal variability using the Last Millennium Reanalysis framework. *Climate of the Past*, **14 (2)**, 157–174, https://doi.org/10.5194/cp-14-157-2018.

Steiger, N. J., G. J. Hakim, E. J. Steig, D. S. Battisti, and G. H. Roe, 2014: Assimilation of Time-Averaged Pseudoproxies for Climate Reconstruction. *Journal of Climate*, **27 (1)**, 426–441, https://doi.org/10.1175/JCLI-D-12-00693.1.

Steiger, N. J., J. E. Smerdon, E. R. Cook, and B. I. Cook, 2018: A reconstruction of global hydroclimate and dynamical variables over the common era. *Scientific Data*, **5 (1)**, 180 086, https://doi.org/10.1038/sdata.2018.86.

Stevenson, S., B. L. Otto-Bliesner, E. C. Brady, J. Nusbaumer, C. Tabor, R. Tomas, D. C. Noone, and Z. Liu, 2019: Volcanic Eruption Signatures in the Isotope-Enabled Last Millennium Ensemble. *Paleoceanography and Paleoclimatology*, **0 (0)**, https://doi.org/10.1029/2019PA003625.

Tardif, R., and Coauthors, 2019: Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling. *Climate of the Past*, **15 (4)**, 1251–1273, https://doi.org/10.5194/cp-15-1251-2019.

Tejedor, E., N. J. Steiger, J. E. Smerdon, R. Serrano-Notivoli, and M. Vuille, 2021: Global hydroclimatic response to tropical volcanic eruptions over the last millennium. *Proceedings of the National Academy of Sciences*, **118 (12)**, e2019145 118, https://doi.org/10.1073/pnas.2019145118.

Thomson, D. J., 1982: Spectrum estimation and harmonic analysis. *Proc. IEEE*, **70(9)**, 1055–1096.

Tierney, J. E., J. Zhu, J. King, S. B. Malevich, G. J. Hakim, and C. J. Poulsen, 2020: Glacial cooling and climate sensitivity revisited. *Nature*, **584 (7822)**, 569–573, https://doi.org/10.1038/s41586-020-2617-x.

Tingley, M. P., and P. Huybers, 2010a: A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part 1: Development and applications to paleoclimate reconstruction problems. *J. Clim.*, **23**, 2759–2781, https://doi.org/10.1175/2009JCLI3016.1.

Tingley, M. P., and P. Huybers, 2010b: A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part 2: Comparison with the Regularized Expectation-Maximization Algorithm. *J. Clim.*, **23**, 2782–2800, https://doi.org/2009JCLI3016.1.

Tingley, M. P., and P. Huybers, 2013: Recent temperature extremes at high northern latitudes unprecedented in the past 600 years. *Nature*, **496 (7444)**, 201–205, https://doi.org/10.1038/nature11969.

Valler, V., J. Franke, Y. Brugnara, and S. Brönnimann, 2022: An updated global atmospheric paleo-reanalysis covering the last 400 years. *Geoscience Data Journal*, **9 (1)**, 89–107, https://doi.org/10.1002/gdj3.121.

Watanabe, S., and Coauthors, 2011: MIROC-ESM 2010: Model description and basic results of CMIP5-20c3m experiments. *Geoscientific Model Development*, **4 (4)**, 845–872, https://doi.org/10.5194/gmd-4-845-2011.

Widmann, M., H. Goosse, G. van der Schrier, R. Schnur, and J. Barkmeijer, 2010: Using data assimilation to study extratropical northern hemisphere climate over the last millennium. *Climate of the Past*, **6**, 627–644.

Wikle, C. K., and L. M. Berliner, 2007: A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, **230 (1–2)**, 1 – 16, https://doi.org/10.1016/j.physd.2006.09.017.

Wu, T., and Coauthors, 2014: An overview of BCC climate system model development and application for climate change studies. *Journal of Meteorological Research*, **28 (1)**, 34–56, https://doi.org/10.1007/s13351-014-3041-7.

Yi, B., F. Viens, B. Law, and Z. Li, 2015: Dynamic portfolio selection with mispricing and model ambiguity. *Annals of Finance*, **11 (1)**, 37–75, https://doi.org/10.1007/s10436-014-0252-y.

Zhu, F., J. Emile-Geay, K. J. Anchukaitis, G. J. Hakim, A. T. Wittenberg, M. S. Morales, M. Toohey, and J. King, 2022: A re-appraisal of the ENSO response to volcanism with paleoclimate data assimilation. *Nature Communications*, **13 (1)**, 747, https://doi.org/10.1038/s41467-022-28210-1.

Zhu, F., J. Emile-Geay, G. J. Hakim, J. King, and K. J. Anchukaitis, 2020: Resolving the Differences in the Simulated and Reconstructed Temperature Response to Volcanism. *Geophysical Research Letters*, **47 (8)**, e2019GL086 908, https://doi.org/10.1029/2019GL086908.

Zhu, F., and Coauthors, 2019: Climate models can correctly simulate the continuum of global-average temperature variability. *Proceedings of the National Academy of Sciences*, **116 (18)**, 8728, https://doi.org/10.1073/pnas.1809959116.