Revisiting Online and Offline Data Assimilation Comparison for Paleoclimate Reconstruction: An Idealized OSSE Study

Atsushi Okazaki\textsuperscript{1,2,3}, Takemasa Miyoshi\textsuperscript{1,4,5,6,7,8}, Kei Yoshimura\textsuperscript{3}, Steven J. Greybush\textsuperscript{4,9}, and Fuqing Zhang\textsuperscript{3}

\textsuperscript{1}Department of Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, PA, USA, \textsuperscript{2}RIKEN Center for Computational Science, Kobe, Japan, \textsuperscript{3}Institute of Industrial Science, The University of Tokyo, Kashiwa, Japan, \textsuperscript{4}RIKEN Interdisciplinary Theoretical and Mathematical Sciences Program, Kobe, Japan, \textsuperscript{5}RIKEN Cluster for Pioneering Research, Kobe, Japan, \textsuperscript{6}Department of Atmospheric and Oceanic Science, University of Maryland, College Park, \textsuperscript{7}Graduate School of Science, Kyoto University, Kyoto, Kyoto, Japan, \textsuperscript{8}Application Laboratory, Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan

Abstract

Data assimilation (DA) has been applied to estimate the time-mean state, such as annual mean surface temperature for paleoclimate reconstruction. There are two types of DA for this purpose: online-DA and offline-DA. The online-DA estimates both time-mean states (analyses) and initial conditions for subsequent DA cycles, while the offline-DA only estimates the time-mean analyses. If there is sufficiently long predictability in the system of interest compared to the temporal resolution of the observations, online-DA is expected to outperform offline-DA by utilizing information in the initial conditions. However, previous studies failed to show the superiority of online-DA when time-averaged observations are assimilated, and the reason has not been investigated thoroughly. This study compares online-DA and offline-DA and investigates the relation to the predictability using an intermediate complexity general circulation model with perfect-model observing system simulation experiments. The result shows that the online-DA outperforms offline-DA when the length of predictability is longer than the averaging time of the observations. We also found that the longer the predictability, the more skillful the online-DA. Here, the ocean plays a crucial role in extending predictability, which helps online-DA to outperform offline-DA. Interestingly, the observations of near-surface air temperature over land are highly valuable to update the ocean variables in the analysis steps, suggesting the importance of using cross-domain covariance information between the atmosphere and the ocean when online-DA is applied to reconstruct paleoclimate.

1. Introduction

Knowledge of past climate conditions is crucial to understanding the climate system. Recently, data assimilation (DA) has been applied to reconstruct paleoclimate of the last millennium (e.g., Franke et al., 2017; Goosse et al., 2010, 2012a; Hakim et al., 2016; Steiger et al., 2018; Tardif et al., 2019) and ages deeper into the past (e.g., Kurahashi-Nakamura et al., 2017; Mathiot et al., 2013; Renssen et al., 2015; Tierney et al., 2020). DA estimates the most likely state (analysis) by combining model simulations (background) and observations. DA has long been used for numerical weather prediction (NWP) and is a well-established method (e.g., Houtekamer and Zhang, 2016; Kalnay, 2003; and references therein). The purpose of DA in NWP is to provide the optimal initial conditions for subsequent forecasts, where the analyses represent instantaneous values and will be used as initial conditions for subsequent DA cycles. On the other hand, DA has been used differently in paleoclimate reconstruction, where DA is used to estimate time-mean states. Here, for paleoclimate reconstruction, climate proxies such as the physical and chemical properties of ice cores, corals, trees, speleothems, etc., which generally represent surface variables (e.g., surface temperature), serve as the observations. Those data represent time-averaged values (typically annual or longer), while the observations for NWP represent instantaneous values. Several DA methods have been proposed to assimilate such time-averaged observations effectively. Although there are a few methods based on nudging (von Storch et al., 2000), forcing singular vectors (van der Schrier and Barkmeijer, 2005), and the variational method (Kurahashi-Nakamura et al., 2017), this study focuses on ensemble-based ones. The first study that proposes
an ensemble-based method for paleoclimate reconstruction is Dirren and Hakim (2005): they proposed an Ensemble Kalman Filter (EnKF) based method that updates only the time-averaged component of state variables. Subsequently, Annan and Hargreaves (2012) and Goosse et al. (2012a, 2012b) applied a particle filter (van Leeuwen, 2009). Those methods estimate not only time-mean states but also initial conditions for the next DA cycles as in NWP and are categorized as online-DA methods. Huntley and Hakim (2010) showed the capability of online-DA when assimilating time-averaged data using a quasi-geostrophic model (QG-model) (Hakim, 2000). They also showed that online-DA's skill decreases with increasing observation averaging time, and the advantage of the online-DA vanishes and reduces to that of non-initialized simulation when assimilating observations whose averaging time is longer than the predictability of the system. Pendergrass et al. (2012) used a QG-model coupled with a slab ocean and showed that the skill of the online-DA increases as ocean depth increases.

Another method to estimate time-averaged states is known as “offline-DA” (Bhend et al., 2012; Goosse et al., 2006; Steiger et al., 2014). In offline-DA, only time-averaged states are analyzed, but the initial conditions for the next DA cycles are not. Hence, the forecasts are nothing more than free runs. The idea behind the method is that the observations are temporally too sparse to constrain the model. As mentioned above, most typically, the observations in the paleoclimate era represent an annual mean or longer. In other words, they are available once a year at most. On the other hand, the predictability of the atmosphere is roughly two weeks (e.g., Lorenz, 1963, 1969). Therefore, the information contained in initial conditions will be lost long before the end of each DA cycle, no matter how close the initial conditions are to the truth. Therefore, at least for the studies using only atmospheric general circulation models (AGCMs), there is no point in analyzing initial conditions and predicting from them for the subsequent DA cycles. There are two types of offline-DA; transient offline-DA and stationary offline-DA. The background in transient offline-DA consists of an ensemble run for the same time period as the temporal resolution of the observations (Bhend et al., 2012; Franke et al., 2017; Goosse et al., 2006). On the other hand, the background in stationary offline-DA consists of a single run and is the same at all the analysis steps (Hakim et al., 2016; Steiger et al., 2014; Tardif et al., 2019). The idea of stationary offline-DA is similar to that of ensemble optimal interpolation (EnOI) (Oke et al., 2002, 2005). The stationary offline-DA is computationally efficient as it does not require an ensemble run. Steiger et al. (2014) showed that the method outperforms the conventional climate field reconstruction method based on principal component analysis in reconstructing the surface temperature field, while its relative performance to transient offline-DA is still open question.

Which DA method is favorable for paleoclimate reconstruction, online-DA or offline-DA? In terms of skill in estimating time-mean states, online-DA should be better than offline-DA if the system of interest has long enough predictability compared to the averaging time that observations represent since online-DA can use better initial conditions and flow-dependent covariances in the case of EnKF-based methods. In terms of computational cost, offline-DA is preferable because it can use existing model simulations as background model forecasts. In contrast, ensemble simulations must be newly conducted for online-DA, which requires tremendous computational cost considering the period it must cover (i.e., hundreds to a thousand years for the last millennium climate reconstruction). To efficiently use the information in initial conditions with online-DA, models containing slowly changing components, such as atmosphere-ocean coupled general circulation models (CGCMs), must be used. However, ensemble-based online-DA with a CGCM is prohibitively expensive. Due to this limitation, online-DA has been applied together with intermediate complexity models to assimilate decadal mean data (e.g., Goosse et al., 2010, 2012a) or with a stochastic model to assimilate annual mean data (Perkins and Hakim, 2017, 2021), and few studies have applied online-DA with a CGCM to assimilate annual mean observations.

Several studies compared the skill of online-DA to that of offline-DA. Matsikaris et al. (2015) used a CGCM known as Max Planck Institute for Meteorology Earth System Model (MPI-ESM) to assimilate continental-scale decadal mean data with an ensemble-based method proposed by Goosse et al. (2006) that selects a member that fits the observations best. Acevedo et al. (2017) used an intermediate complexity AGCM known as SPEEDY (Kucharski et al., 2006; Molteni, 2003) coupled with a slab ocean model to assimilate annual mean data with the EnKF-based method proposed by Dirren and Hakim (2005). However, neither of the studies showed the superiority of online-DA over offline-DA despite the fact that both studies included a slowly changing component. They suspect this may be because the predictability is not long.
enough or the initial conditions are poorly constrained by DA. Owing to the comparable skills, computational efficiency, and less complexity, nowadays the offline-DA has been widely used for paleoclimate applications (e.g., Dee et al., 2016; Franke et al., 2017; Okazaki and Yoshimura, 2017; Steiger et al., 2018; Tardif et al., 2014, 2015; 2019). However, considering that plenty of studies show that the predictability of the surface temperatures, which most proxies represent, are longer than annual (e.g., Collins, 2002; Doblas-Reyes et al., 2013), online-DA must be better than offline-DA when models that have slowly changing components, such as CGCMs, are used. If so, why did the previous studies fail to show the superiority of online-DA? Is this because of the lack of predictability in the models used and/or a poor representation of the initial condition?

This study investigates online-DA’s capability relative to offline-DA together with its relevance to predictability using the SPEEDY coupled with a slab ocean model within a perfect model observing system simulation experiment (OSSE) framework. Our focus is on the last millennium climate reconstruction, where a relatively large number of proxies whose resolution is shorter than annual are available. The SPEEDY can represent realistic physics with low computational cost, which makes it perfectly suited for the feasibility study. It should be worthwhile to investigate this issue with ensemble-based online-DA with a CGCM being a feasible choice in the near future as computation power steadily increases. We also investigate for what length of averaging period online-DA outperforms offline-DA for time-averaged observations (e.g., paleoclimate observations). It is also worth answering such a question since monthly mean observations are available (e.g., for stable water isotope ratios in corals as in Grottoli, 2006). Finally, we will discuss implications to the actual case based on the results of the idealized experiments.

This study involves a CGCM, data assimilation, and hence atmosphere-ocean coupled-DA (CDA). There are two types of CDA; weakly coupled-DA (WCDA) and strongly coupled-DA (SCDA) (Penny and Hamill, 2017 and references therein). WCDA uses a coupled model to generate the background but updates the analysis separately for each component. Namely, atmospheric (oceanic) observations are used to update only atmospheric (oceanic) states. On the other hand, atmospheric (oceanic) observations are used to update both atmosphere and ocean states using the intra-domain and cross-domain error covariance in SCDA. SCDA approach has desirable aspects compared to WCDA; it can use a larger number of observations to update analysis, which reduces uncertainty in the analysis and achieves more self-consistent analysis by reducing the initialization shocks that WCDA can cause. Indeed, several studies showed the superiority of SCDA (e.g., Sluka et al., 2016). However, SCDA does not always produce better results (e.g., Han et al., 2013), and a valid method for SCDA has yet to be established (Penny and Hamill, 2017; WMO, 2017). With a series of CDA experiments conducted in this study, this study may also have implications for CDA and prediction on various time scales beyond paleoclimate DA (e.g., NWP, subseasonal-to-seasonal prediction, decadal prediction).

This paper is structured as follows. Section 2 describes the method to assimilate time-averaged observations and to measure predictability. Section 3 presents the results and a discussion. The summary follows in Section 4.

2. Methods

2.1. Model

The SPEEDY model is an intermediate complexity AGCM (Kucharski et al., 2006; Molteni, 2003). SPEEDY is a hydrostatic model with a primitive equation dynamics and simplified parameterizations that include essential processes such as convection, large-scale condensation, longwave and shortwave radiation, and boundary layer turbulence. SPEEDY has been used to study atmospheric predictability (e.g., Abid et al., 2015; Bahaga et al., 2015; Ehsan et al., 2013) and has been used as a testbed for DA studies first by Miyoshi (2005) and a number of follow-on studies (e.g., Amiezua et al., 2014; Greybush et al., 2011; Hatfield et al., 2018; Kalnay et al., 2007; Kondo & Miyoshi, 2016, 2019; Kotsuki et al., 2020; Li et al., 2009; Miyoshi, 2011; Miyoshi et al., 2014). We choose SPEEDY for its ability to represent realistic physics with a low computational cost. Though SPEEDY does not consider a diurnal cycle, which makes it difficult to accurately represent convection and sub-grid-scale vertical heat transport (Molteni, 2003), this will not be a significant issue for our study with an OSSE framework. SPEEDY has a horizontal resolution of T30 (approximately 400 km
on the equator) and eight vertical levels. SPEEDY can be coupled with a slab ocean model as an optional facility, which does not consider advection or convection but exchanges heat with the atmosphere at the ocean surface. The prognostic variable of the slab ocean model is only sea surface temperature (SST), and the differential equation is given by

$$\Delta SST_{t+1} = \frac{\tau_{ocn}}{\tau_{ocn} + \delta t} \Delta SST_t + \frac{\delta t}{d_{ocn} C_{ocn}} \Delta F_t$$

(1)

where $\Delta SST_t$ and $\Delta F_t$ are the SST and the net heat flux anomaly from the climatological mean at time $t$, $\tau_{ocn}$ the damping timescale (90 d), $\delta t$ the timestep (1 day), $d_{ocn}$ the depth of the slab ocean (50 m), and $C_{ocn}$ the heat capacity ($4.18 \times 10^6$ JK$^{-1}$m$^{-3}$) (http://users.ictp.it/~kucharsk/speedy_description/km_ver40_appendixA.pdf). The net heat flux is given by a sum of shortwave and longwave radiation and sensible heat and moisture fluxes. Here, the sensible heat flux is a function of SST and near-surface air temperature, and the moisture flux is a function of SST and near-surface specific humidity. Through the heat exchange, the ocean affects the evolution of the atmosphere. Since the ocean has a large heat capacity and hence slow variability, it serves as the source of long-term memory for the atmosphere. The length of the ocean memory can be tuned by changing $\tau_{ocn}$ and/or $d_{ocn}$. A simple illustrative example would be the case with infinitely large $\tau_{ocn}$ and/or $d_{ocn}$: in this example, the coefficients of $\Delta SST_t$ and $\Delta F_t$ in Equation 1 converge to one and zero, respectively, bringing the system to a steady state where the initial condition persists forever ($\Delta SST_{t+1} = \Delta SST_t = \Delta SST_{t-1} = \ldots = \Delta SST_{t-\infty}$). Therefore, the system has eternal predictability. We choose to couple SPEEDY with the slab ocean model to elongate the predictability of the system so that it can effectively assimilate time-averaged observations whose averaging time is beyond the predictability of the atmosphere and to investigate the response of online-DA and offline-DA skills to the length of predictability.

### 2.2. Data Assimilation Method

This study uses the time-average update (TAU) method proposed by Dirren and Hakim (2005). The method updates only the time-averaged component of the model state and leaves the deviation from the time-mean unchanged. The time-averaged component and the deviation represent the low- and high-frequency components of the system, respectively. Let $x_n^b$ be an $N$-dimensional background model forecast at time $t_n$. The TAU first decomposes $x_n^b$ into the time-averaged component $\overline{x}_{t_n}$ and the deviation from it $x_{t_n}^b$:

$$x_n^b = \overline{x}_{t_n} + x_{t_n}^b$$

(2)

$$\overline{x}_{t_n} = \frac{1}{\tau} \sum_{t = t_n-1}^{t_n} x_t$$

(3)

where $\tau = t_n - t_{n-1}$ defines the time-averaging length. For the time being, the time notations will be omitted for simplicity. After the decomposition, the TAU updates the time-averaged component. This study uses the Local Ensemble Transform Kalman Filter [LETKF] (Hunt et al., 2007), a variant of the EnKF (Evensen, 1994), to update the time-averaged component. Here, we consider an ensemble of $m$ members. In LETKF, the analysis ensemble mean and the ensemble perturbations are given by:

$$\overline{\{x^a\}} = \overline{\{x\}} + \overline{X^a P^o (HX^a)^T} R^{-1} (y^o - \tau \overline{\{x^b\}})$$

(4)

$$\overline{X^a} = \overline{X^b} \sqrt{m - 1} P^o^{-1/2}$$

(5)

where $\overline{\{x^a,b\}}$ denotes the ensemble mean and $X^{a,b}$ the ensemble perturbation matrix, and the superscripts $a$ and $b$ denote analysis and background, respectively. Here $X^{a,b}$ is the $N \times m$ matrix whose $i$th column is $x^{a,b}_i = \overline{x^{a,b}}$, where superscript $(o)$ denotes that it is the $t$th member of the ensemble ($i = \{1,2,\ldots,m\}$). The notation $y^o$ is the observation vector of length $p$, $R$ the observation error covariance matrix whose size is $p \times p$, $H$ the observation operator that converts the model state to the observation equivalent quantity,
The linearized observation operator whose size is $p \times N$, and $P_{a\pm}$ the covariance matrix in the ensemble space given by

$$P_{a\pm} = (m - 1)I + \left(\mathbf{HX}_{\pm}^T\right)^{-1} \mathbf{R}^{-1} \left(\mathbf{HX}_{\pm}^T\right)^{-1}.$$  

(6)

Then the $i$th member of time-averaged analysis $\overline{x^d}$ is obtained by adding $\overline{x^d}$ to the $i$th column of $\overline{X^d}$. Finally, the time-averaged analysis and the background anomaly are combined to obtain the full analysis field at time $t_n$:

$$x_{a^i}^n = \overline{x^n_{a\pm-1}} + x_{b^i}^n.$$  

(7)

In online-DA, $x_{a^i}^n$ is used as the initial condition for the following forecast to obtain $x_{b^i+1}^n$ while $x_{b^i}^n$ is used in the transient offline-DA. In stationary offline-DA, model integration is not involved, and $x^b$ is the same at all time steps ($x_{b^n}^n = x_{b^{n-1}}^n = \ldots = x_{b^1}^1$).

2.3. Experimental Design

We perform a series of OSSEs with the perfect model assumption where model errors are excluded. The nature run (i.e., the truth) is created by integrating SPEEDY for 100 years from a date arbitrarily denoted Jan 1st, 1900, after a 400-year spin-up period, where the model starts from the atmosphere at rest (wind speed is zero everywhere). The nature run is externally forced by solar and greenhouse gases, but the irradiance and the concentrations are both constant over time. Note that there are seasonal cycles as the declination angle changes in a year.

We created seven sets of observations. Each set of observations differs in the length of averaging time ($\tau$). The values of $\tau$ considered are 1, 3, 10 days, 1, 3, 6, and 12 months. Surface temperature data are drawn from the nature run, and the observations are generated in a way that mimics the real proxy network for the last millennium that consists of the composite of Mann et al. (2008), PAGES 2k consortium (2013), and Okazaki and Yoshimura (2017). When an observation is on the ocean grid in SPEEDY, it is assumed to represent sea surface temperature (SST). Otherwise, the observations represent surface air temperature (SAT). Therefore, SAT (SST) observations are all over land (ocean). This results in 413 SAT and 34 SST observation points (Figure 1a).

Each observation is assumed to have an error of 1/2 of the climatological variance for an aver-
aging time of the observations (Figures 1b–1h). Note that the seasonality is subtracted from each time series before calculating the variance when the averaging time is shorter than 12 months.

Data assimilation experiments are performed with online-DA (ONLINE) and offline-DA (OFFLINE) and without DA (NODA) (Table 1). For experiments with offline-DA, transient offline-DA is used in this study. Each experiment except for NODA consists of seven sub experiments, each of which assimilates one set of observations. That is, the sub experiment 1, 2, ..., 7 assimilates observations with τ of 1, 3, 10 days, 1, 3, 6, and 12 months, respectively. Note that the analysis interval of DA-cycles is the same as the averaging time τ of the observation. Each experiment consists of one ensemble integration of 100 DA-cycles starting from January 1st, 1900. Thus, the number of observations is constant regardless of the averaging time. For all the experiments, the ensemble size is fixed to 30. We determined the ensemble size by considering that at this time, the computation is too costly to run a larger ensemble with a full CGCM when online-DA is applied to the real problem. For both ONLINE and OFFLINE, the atmospheric (oceanic) observations are used to update both atmosphere and ocean states (i.e., SCDA). The initial conditions for all the experiments are identical and are drawn from dates in subsequent years of the nature run (i.e., Jan 1st 2000, 2001,..., 2029).

Another set of experiments similar to ONLINE, OFFLINE, and NODA but with \( \tau_{OCN} \) of 360 days is performed to investigate the response of skills in online-DA and offline-DA to the length of predictability (Table 1). After changing the parameter value, we re-generated the nature run and the observations in the same way and repeated all the DA experiments. The experiments with \( \tau_{OCN} \) of 360 days will be referred to as ONLINE_LONG, OFFLINE_LONG, and NODA_LONG. Note that the initial conditions used in the experiments are different from those used in ONLINE, OFFLINE, and NODA as they are drawn from the respective nature run.

Another four experiments with online-DA are performed to investigate the relative impact of the atmosphere and ocean in ONLINE (Table 1): ONLINE_AA, ONLINE_OO, ONLINE_AO, and ONLINE_OA. The ONLINE_AA (ONLINE_OO) assimilates only SAT (SST) to update atmospheric (oceanic) variables, and the oceanic (atmospheric) variables are not updated. Namely, only information within each domain is used, but cross-domain information is not in ONLINE_AA and ONLINE_OO in the analysis steps. Conversely, only cross-domain information is used in ONLINE_OA and ONLINE_AO: the ONLINE_OA (ONLINE_AO) assimilates only SAT (SST) to update oceanic (atmospheric) variables. The improvement in SST in ONLINE_AA and ONLINE_OA compared to NODA is due to the interaction between the atmosphere and ocean in the forecast steps. On the other hand, the improvement in SST in ONLINE_AO and ONLINE_OO compared

| Experiment        | \( \tau_{OCN} \) (d) | DA method    | Observation | Update variables |
|-------------------|-----------------------|--------------|-------------|------------------|
| NODA              | 90                    | -            | -           | -                |
| OFFLINE           | 90                    | Offline-DA   | SAT, SST    | T, U, V, Q, Ps, SST, LST |
| ONLINE            | 90                    | Online-DA    | SAT, SST    | T, U, V, Q, Ps, SST, LST |
| ONLINE_LONG       | 360                   | -            | -           | -                |
| OFFLINE_LONG      | 360                   | Offline-DA   | SAT, SST    | T, U, V, Q, Ps, SST, LST |
| ONLINE_LONG       | 360                   | Online-DA    | SAT, SST    | T, U, V, Q, Ps, SST, LST |
| ONLINE_AA         | 90                    | Online-DA    | SAT         | T, U, V, Q, Ps, LST |
| ONLINE_OO         | 90                    | Online-DA    | SST         | SST              |
| ONLINE_OA         | 90                    | Online-DA    | SST         | T, U, V, Q, Ps, LST |
| ONLINE_OA         | 90                    | Online-DA    | SAT         | SST              |

Note. Each experiment consists of seven sub experiments. Each sub experiment assimilates one set of observations whose averaging time is 1, 3, 10 days, 1, 3, 6, and 12 months. The variables denoted as T, U, V, Q, Ps, SST, and LST represent temperature, zonal wind, meridional wind, specific humidity, surface pressure, sea surface temperature, and land surface temperature.
to NODA is due to updating the oceanic variables in the analysis steps. Note that the initial conditions used in the experiments are identical to those used in ONLINE, OFFLINE, and NODA.

For experiments using online-DA, the relaxation-to-prior spread method (RTPS) (Whitaker and Hamill, 2012) is applied. In addition, covariance localization using a Gaspari-Cohn function (Gaspari & Cohn, 1999) is used for online-DA and offline-DA experiments. Combination of the relaxation parameter \( \alpha = \{0.0, 0.6, 0.9\} \) in Equation 2 in Whitaker and Hamill, 2012 and localization scale (radius of influence \( ROI = \{2000km, 4000km, 6000km, 8000km\} \) are manually tuned for each experiment so that the analysis root-mean-square error (RMSE) of the SAT is minimized. The experiments with an optimal combination will be shown elsewhere.

We use RMSEs to measure the skills of each DA method. We compute RMSEs for the ensemble mean of the time-averaged component \( \langle x \rangle \). Only sea-ice free areas are used in calculating the global mean RMSEs for SST as SST at sea-ice covered area is output as the average of SAT and SST weighted by sea-ice covered and sea-ice free areas (e.g., SST is output as SAT if the ocean is fully covered by ice).

### 2.4. Predictability

We measure the predictability of SAT and SST with ensemble forecasts and anomaly correlation coefficient (ACC). The ensemble forecasts are initialized in four different ways. In the first case, the initial conditions are identical to those for the nature run but with small independent perturbations drawn from the normal distribution with zero mean and standard deviation of \( 10^{-6} \). In the second one, the initial conditions are drawn from the analysis ensemble members of ONLINE. In the third one, the initial conditions are drawn from dates in subsequent years of the nature run, and the time-averaged component of the initial conditions is replaced by that of the nature run and leaves the deviation from the time-averaged component unchanged. In the last case, the initial conditions are drawn from the first guess ensemble members of OFFLINE. The first one measures the upper bound of the predictability of SPEEDY. Although a larger ensemble size than that used in this study may result in longer predictability by increasing the signal-to-noise ratio (e.g., Scaife et al., 2014), hereafter we refer to the predictability as "intrinsic predictability" (Lorenz, 2006; Melhauser and Zhang, 2012; Zhang et al., 2006) for convenience. The second one measures the predictability of SPEEDY assimilating time-averaged observations with TAU, which will be referred to as “practical predictability” (Lorenz, 2006; Melhauser and Zhang, 2012; Zhang et al., 2006). In this case, errors are in both low- and high-frequency components. The former is from the analysis error, and the latter is because TAU does not constrain the high-frequency component (c.f., Section 2.2). The third one measures predictability when there is no error in the low-frequency component, but there is an error in the high-frequency component. Thus, it reveals the upper limit of the predictability when TAU is used to initialize the ensemble forecasts. Hereafter, we refer to the predictability as "potential predictability.” The last one serves as the baseline where no initialization is conducted. Hereafter, we refer to the predictability as NOINT. Intrinsic predictability is “intrinsic” to the system of interest. On the other hand, potential predictability is affected by DA method (i.e., TAU in this study), and practical predictability is affected by DA experiment settings (e.g., localization, covariance inflation, etc.) and the number of and errors in observation as well as DA method. Therefore, we can infer what hampers the practical predictability by comparing these predictabilities. For all the forecasts, the ensemble size is 30, and the forecast length is five years. A set of the ensemble forecasts is repeated 60 times with different starting dates of the forecasts. ACC is calculated for the time-averaged component of SAT and SST using the ensemble mean of the forecasts and the nature run. For instance, when calculating ACC for 1 month means, samples used to calculate ACC are also monthly mean. The anomaly is calculated with regard to the climatological seasonality for \( \tau \) shorter than 12 months. When \( \tau \) is 12 months, the anomaly is calculated with regard to the climatological mean. All the ensemble forecasts are performed with \( r_{OCE} \) of 90 and 360 days. In this study, a variable with ACC significantly larger than zero at a 95% confidence level is assumed to be predictable as in the previous climate prediction studies (e.g., Collins, 2002), where a one-sided Student t-test is used for the statistical test.
3. Results and Discussion

3.1. Case With $\tau_{OCN}$ of 90 days

This section compares online-DA and offline-DA with $\tau_{OCN}$ of 90 days. The global mean RMSEs normalized by that of NODA are shown in Figure 2. For SAT, the RMSEs of ONLINE are smaller than those of OFFLINE when $\tau$ is shorter than or equal to 10 days ($p < 0.05$). Within this range of $\tau$, the RMSE for ONLINE is smaller than OFFLINE by 23% (1 day) to 3% (10 days). With $\tau \geq 1$ month, the RMSEs for ONLINE are indistinguishable from those for OFFLINE. For SST, the RMSEs of ONLINE are smaller than those of OFFLINE with $\tau \leq 3$ months ($p < 0.05$) and are indistinguishable from those of OFFLINE with $\tau \geq 6$ months. The result indicates that ONLINE outperforms OFFLINE if $\tau$ is sufficiently short.

We expected that the skills of ONLINE and OFFLINE decrease as $\tau$ lengthens, as shown in Pendergrass et al. (2012). Interestingly, however, the skill of OFFLINE increases along with $\tau$ for SST. Similarly, the skill of ONLINE for SST does not decrease as $\tau$ lengthens with the skill with $\tau$ of 3 days being remarkably worse. These results are owing to a relatively small ensemble size to suppress spurious correlations between SST and atmospheric variables including SAT when $\tau$ is short. Section 3.4 will investigate the issue in detail.

Figure 2. Global mean root mean square error (RMSE) difference (XXX-NODA) for (a) Surface air temperature and (b) Sea surface temperature, where XXX = {OFFLINE, ONLINE, ONLINE_AA, ONLINE_OA, ONLINE_OA, ONLINE_OO}. RMSEs are normalized by the global mean RMSE of NODA for each averaging time $\tau$. Asterisk (*) above red bars denotes RMSE of ONLINE is statistically smaller than OFFLINE at the significance level of 0.05.
3.2. Predictability

Figure 3 shows the global mean ACC for SAT and SST. We determine whether a variable is predictable or not based on ACC. The time-averaged SAT and SST are intrinsically predictable for the length of $\tau$ for all the $\tau$. For example, the 1-month mean SAT is predictable one month after the forecasts have started. Here, 1-month lead time means the average of 0-1-month lead time. The 1, 3, and 10-day mean SAT are practically predictable during at least 1, 3, and 10 days after the forecasts are launched. The time-averaged SST except for the 12-month mean are practically predictable for the length of $\tau$.

Correspondences can be found among practical predictabilities and the skills of ONLINE relative to OFFLINE (Figures 2 and 3). That is, when the time-averaged state is predictable for the period of averaging time, ONLINE outperforms OFFLINE. For instance, 1-month mean SST is predictable at a lead time of 1 month, and the online-DA outperforms offline-DA when assimilating 1-month mean SST. Conversely, the skill of ONLINE is indistinguishable from that of OFFLINE when the time-averaged state is not predictable for the period of $\tau$ after the forecast is launched. The correspondences suggest that the practical predictability controls whether online-DA outperforms offline-DA or not. Although the relationship between the predictability and online-DA skill relative to offline-DA can be anticipated based on the theory, this study is the first to show the relationship for paleoclimate reconstruction applications.

The finding confirms the importance of practical predictability for online-DA to outperform offline-DA. The practical predictability can be further extended by improving DA method, settings, and/or adding more observations and reducing the errors in them. Now, there are significant gaps between the intrinsic and practical predictability both for SAT and SST, suggesting significant impacts the DA method (i.e., TAU), the DA experimental settings (e.g., localization, covariance inflation, etc), and the number of and errors in observations have on the predictability. Here, we briefly investigate what limits the predictability by comparing the three types of predictability. The potential predictability reveals the upper bound of the predictability with TAU (c.f., Section 2.4). Therefore, the difference between intrinsic and potential predictability shows the extent to which the DA method limits the capability of online-DA, and the difference between the potential and practical predictability shows the extent to which the DA experiment settings and the quality and quantity of the observations limit the capability of online-DA. Comparing the predictabilities reveals that the loss of practical predictability for SST is mainly due to DA experimental settings when $\tau \leq 10$ days while it should be due to DA method when $\tau$ is 12 months. Therefore, although the capability of online-DA is largely capped by the intrinsic nature of the climate system, it is also limited by the DA method and settings.
This implies further potency of online-DA if one can develop a better DA method, settings, and/or with more observations and smaller errors in them.

### 3.3. Case With $\tau_{OCN}$ of 360 days

Section 3.2 shows that the online-DA is beneficial when there is long enough practical predictability compared to the averaging time of observations. To further support the claim, we examine how the skill of ONLINE relative to OFFLINE changes with the length of predictability. This experiment would help provide implications on how to translate this study’s result into a more realistic case, where the ocean has much longer predictability (e.g., 6–9 years for SAT at annual to subdecadal time scales; Doblas-Reyes et al., 2013) than the model we used in this study (i.e., SPEEDY coupled with a slab ocean). Figure 3 shows the practical predictabilities with $\tau_{ocn}$ of 360 days in red. The ACCs become stronger, and the predictabilities become longer for both SAT and SST. The prolonged practical predictability is apparent in the 6-month mean SST.

Figure 4 shows the root mean square errors (RMSEs) for ONLINE_LONG alongside ONLINE. For SAT, the RMSE remains the same as that of the experiment with $\tau_{ocn}$ of 90 days: with $\tau \leq 10$ days, ONLINE_LONG is better than OFFLINE_LONG, and they are equally skillful with $\tau \geq 1$ month. On the other hand, for SST, the skills of online-DA relative to offline-DA with $\tau_{OCN}$ of 360 days have been improved with $\tau$ of 1, 3, and 6 months, compared to those with $\tau_{OCN}$ of 90 days. For $\tau$ of 1 and 3 months, the RMSE of ONLINE_LONG decreased to 0.88 and 0.91, respectively, which were initially 0.92 and 0.96 in ONLINE. For $\tau$ of 3 months, the significance level also improved from $p < 0.05$ to $p < 0.001$. Furthermore, ONLINE_LONG outperforms OFFLINE_LONG for $\tau$ of 6 months, which was not the case in the experiment with $\tau_{ocn}$ of 90 days. Therefore, the result clearly shows the relationship between the predictability and online-DA skill relative to offline-DA: the longer the predictability, the more beneficial the online-DA.

![Figure 4](image-url). Global mean root mean square (RMSE) for (a) Surface air temperature and (b) Sea surface temperature for each averaging time. Red and green boxes indicate ONLINE ($\tau_{ocn} = 90$ days) and ONLINE_LONG ($\tau_{ocn} = 360$ days). The RMSEs are normalized by the global mean RMSE of the corresponding offline-DA experiment. The centerline in the boxes, the bottom and top of the boxes show median, 25 and 75 percentiles, and whiskers show the range of 1 and 99 percentile of the RMSE. Asterisks *, **, and *** under the boxes show that the RMSE of ONLINE (ONLINE_LONG) is significantly smaller than OFFLINE (OFFLINE_LONG) at the significance level of 0.05, 0.01, and 0.001, respectively.
In Section 3.1, we mentioned that the global mean RMSE of OFFLINE for SST increases along with $\tau$, and that of SST in ONLINE has a local maximum when $\tau$ is 3 days, as opposed to our intuition (Figure 2). The spatial distribution of RMSE would help understand the unexpected results. For OFFLINE with $\tau$ of 1 day, the areas where $\Delta$RMSE (OFFLINE-NODA) is negative are confined around SST observations while the $\Delta$RMSE is positive around SAT observations (Figure 5a). Along with $\tau$, the areas where $\Delta$RMSE is negative around SST observations spread, and the areas where $\Delta$RMSE is positive disappear and turn into negative even around SAT observations, suggesting that assimilating SAT is detrimental for updating SST when $\tau$ is short, while it is beneficial when $\tau$ is long. Figure 6 shows the correlations between SAT and SST for each $\tau$ as a function of their distance. For that, we computed single-point correlations between SAT at each model grid point and SST fields using the nature run. The seasonality is subtracted before the computation. The sample size is 100 for all $\tau$. Figure 6 shows that the correlation between SAT and SST is mostly weak and not statistically significant except for a very close distance (500 km) when $\tau \leq 3$ days. With longer $\tau$, the correlations are stronger. This suggests that the degradation in the skill of OFFLINE with short $\tau$ should be due
to the weak correlation between SAT and SST and the insufficient ensemble size to suppress the spurious correlations. Indeed, the detrimental impact of DA on SST is largely mitigated in an offline-DA experiment with a 100-member ensemble (not shown).

For ONLINE, the degraded RMSE compared to NODA is notable when \( \tau \leq 3 \) days as in OFFLINE but with a larger amplitude (Figures 5e–5g). As online-DA transfers either “good” or “bad” information to the subsequent DA-cycles, such information is accumulated in ONLINE, resulting in a larger amplitude in RMSE compared to OFFLINE. It is difficult to readily interpret what degrades the skill of SST in ONLINE only from Figure 5 as it is affected both in analysis and forecast steps. To disentangle the cause of the degraded skill in SST, we conducted another four experiments with online-DA: ONLINE_AA, ONLINE_OO, ONLINE_AO, and ONLINE_OA (see Section 2.3 and Table 1). Negative impacts of DA are apparent in ONLINE_AA for \( \tau \leq 3 \) days (Figures 5i and 5j). With increasingly larger \( \tau \), the negative impact decreases, and it becomes positive overall (Figures 5k and 5l). For \( \tau \leq 3 \) days, the RMSEs are larger than NODA at a distance from the SAT observations while they are smaller at around the observations. In other words, the skills are degraded over the middle of the oceans and improved along the coastlines. A similar pattern can be found in RMSE for SAT in ONLINE_AA (not shown), probably due to the suboptimal localization scale and covariance inflation factor for the areas where observations are sparse, suggesting that the atmosphere disturbs the ocean there. ONLINE_AO similarly exhibits the negative impact of DA for \( \tau \) of 1 day (Figure 5q). This is because of the weak correlation between SAT and SST when \( \tau \) is small (Figure 6a). With increasingly larger \( \tau \), the negative impact is softened (Figures 5r and 5s), and eventually, the impact becomes positive globally (Figure 5t) due to the strong correlation between SAT and SST when \( \tau \) is large (Figure 6d). The impact of updating SAT by assimilating SST is marginal for all \( \tau \) (Figures 5m–5p), probably because the number of SST observations is small. The impact of updating SST by assimilating SST is overall positive, resulting in decreased RMSE in ONLINE_OO globally for all \( \tau \) (Figures 5u–5x).

The global mean RMSEs for the experiments with online-DA are summarized in Figure 2. For SAT, regardless of \( \tau \), most of the improvements in ONLINE come from SAT observations to update the atmospheric variables, and updating atmospheric and oceanic variables by assimilating SST and updating oceanic variables by assimilating SAT play minor roles. On the other hand, for SST, each plays a considerable role in ONLINE with updating atmospheric variables by assimilating SAT being the largest for \( \tau \leq 3 \) days, showing that the SST is improved in the forecast steps through heat exchange with the atmosphere. Within the timescales, the impact SAT observations have on directly updating SST was negative due to the weak correlation between SAT and SST. With \( \tau \geq 1 \) month, the benefit of interacting with the atmosphere is small because
the atmospheric predictability is shorter than that of the ocean, and the atmosphere perturbs the ocean. Interestingly, within the timescale, the large portion of the improvement in SST in ONLINE depends on updating SST with SAT observations. This suggests the importance of the atmospheric observations over land for the ocean in these timescales.

To corroborate the importance of the atmospheric observations, we measured the impact of observations using analysis sensitivity to the observation (AS). AS was originally proposed for 4D-Var by Cardinali et al. (2004) and later expanded for EnKF by Liu et al. (2009). See Text S1 for more detail. The global mean of AS is shown in Figure 7 for all the $\tau$. Since the localization scale and covariance inflation coefficient influence the AS, comparison among different averaging times does not make sense. On the other hand, we can understand the relative importance of SST observations to SAT observations from the figure. The figure shows that the total impact of SST observations is larger than that in SAT observations when $\tau$ is 1 day and 3 days. With longer $\tau$, SAT observations have a larger impact than SST observations. On the contrary, the per-observation impact is larger in SST than in SAT for all $\tau$ (Figure 7). This implies that the larger impact of SAT is because of the larger number of SAT observations than SST (413 and 34, respectively). This confirms that the impact of SAT observations is larger than that of SST observations. Although this statement is certainly observation distribution dependent, this is true with the present observation network where most of the observations are SAT. The result indicates that the SAT observations over land are highly valuable especially when considering that most of the observations for the last millennium are over land and represent SAT in reality (Figure 1a), and the correlations between SAT and SST are fairly large at these timescales (Figure 6).

It remains unclear why ONLINE has a local maximum in RMSE at 3 days. According to Dirren and Hakim (2005), TAU is expected to be more accurate when time-averaged and the deviation components are independent. Possibly, SPEEDY might not satisfy this condition when $\tau$ is 3 days, and it resulted in the degraded skill.

### 3.5. Implications to the Real Case

We showed in Section 3.3, the longer the predictability, the more beneficial the online-DA compared to offline-DA when assimilating observations whose temporal resolution is shorter than the length of predictability. Now, in the real case (i.e., real paleoclimate observations and a more realistic CGCM), the temporal resolution of the paleoclimate proxies for the last millennium is most typically annual. On the other hand, the annual mean surface temperature is practically predictable for several years (e.g., Collins, 2002; Chikamoto et al., 2013; Doblas-Reyes et al., 2013). Given that the online-DA is beneficial when the predictability

---

**Figure 7.** Global mean analysis sensitivity of sea surface temperature (SST) to the observations. White and gray bars show the impact of surface air temperature and SST observations, respectively. (a) Total impact and (b) per-observation impact are shown as a function of $\tau$. 
is longer than averaging time (i.e., temporal resolution of the observations), the results imply that the online-DA should outperform offline-DA even in the real case.

We also showed that in the experiment with online-DA with $\tau \geq 1$ month, the skill for SST is largely owing to the SAT observations in updating the oceanic variables in analysis steps. In turn, this is due to the strong correlation between SAT and SST and the fact that the bulk of the observations represent SAT. Figure 6 shows the correlation between SAT and SST in a more sophisticated CMIP5-class CGCM named MIROC5 (Watanabe et al., 2010). Simulations from the preindustrial control experiment are used to compute the correlation (see Taylor et al., 2012 for detailed information on the experimental design). High correlations between SAT and SST with $\tau \geq 1$ month compared to those for shorter timescales can also be found in MIROC5 (Figures 6h–6n). Therefore, it is legitimate to expect that the ocean state is improved with SAT observations even with a more sophisticated CGCM with sufficiently long $\tau$.

Lastly, it is important to note that online-DA should adopt SCDA to use atmospheric observations over land effectively. Otherwise, SST will not be updated by SAT observations in the analysis steps, and the information contained in SAT observations will be lost before it affects the ocean during the model integration because of the loss of predictability in the atmosphere.

4. Summary

This study compares online-DA and offline-DA skills when assimilating time-averaged data using an intermediate complexity AGCM, known as SPEEDY, coupled with a slab ocean model under a perfect model OSSE scenario. We found that online-DA outperforms offline-DA if there is practical predictability longer than the temporal resolution of the observations. We further confirmed the results by changing the length of the predictability of the ocean in the SPEEDY model. We found that the longer the predictability, the more beneficial the online-DA compared to offline-DA. This study shows for the first time the superiority of online-DA using a CGCM when assimilating time-averaged data whose averaging time is on the time scales of the paleoclimate data.

Considering that the typical temporal resolution of the proxies is annual for the last millennium and that annual mean surface temperature is practically predictable for several years (e.g., Collins, 2002; Chikamoto et al., 2013; Doblas-Reyes et al., 2013) in the real case, we can expect that online-DA improves the skill of paleoclimate reconstruction. Furthermore, although the ocean plays a vital role in carrying the information contained in the observations over time, we found the importance of the atmospheric observations over land in updating the ocean variables in the analysis steps. Therefore, the online-DA should adopt atmosphere-ocean strongly coupled-DA (SCDA) to effectively use the information contained in observations.

On the other hand, when assimilating observations whose averaging time is equal to or shorter than 3 days, SCDA resulted in degraded skills for SST. This is due to the limited ensemble size and low signal-to-noise ratio in the background error covariance. The degraded skill when ensemble size is small is commonly seen in SCDA (e.g., Han et al., 2013; Yoshida and Kalnay, 2018). In such a case, localizing the background error covariance is essential to obtain a better analysis. Indeed, the degradation in SST can be mitigated with a tighter localization scale in SCDA (not shown), which indicates that tuning of localization scale for cross-domain covariance or correlation cutoff method as suggested by Yoshida and Kalnay (2018) should be effective. This study also confirms that the time-averaging strategy is effective for SCDA, as Lu et al. (2015a, 2015b) showed.

While this study shows the superiority of online-DA to offline-DA for SST, the benefit of online-DA for the atmospheric variables is found to be limited. In reality, the atmosphere can be influenced by predictable slowly changing SST in the Pacific and the Atlantic, for example, ENSO (Trenberth and Caron, 2000) and Atlantic Multidecadal Oscillation (Knight et al., 2006). One possible reason for the limited impact on the atmosphere is that such internal variabilities are lacking in SPEEDY’s current configuration in which the model is coupled to the slab ocean. It is also unclear how well the variability in a subsurface ocean, where prominent predictability resides (e.g., Chikamoto et al., 2013), can be reconstructed using online-DA assimilating only surface variables, which most proxies represent. In the future, we will investigate the impact of online-DA on the atmosphere and the subsurface ocean using a more sophisticated CGCM.
This study assumed a perfect model for all the experiments where the model error is excluded. However, in reality, a model error has been a major issue in atmosphere-ocean coupled-DA and decadal prediction (Meehl et al., 2009, 2014; WMO, 2017 and references therein). It is also an important task to develop an effective way to reduce the biases to realize online-DA with a CGCM.

Data Availability Statement

CMIP5 outputs are available at https://esgf-node.llnl.gov/projects/cmip5/. All the programs except for those of SPEEDY are available at https://github.com/ats-okazaki/Okazaki-et-al_2021_JGR-A.
Hatfield, S. E., Düben, P. D., Chantry, M., Kondo, K., Miyoshi, T., & Palmer, T. N. (2018). Choosing the optimal numerical precision for data assimilation in the presence of model error. *Journal of Advances in Modeling Earth Systems, 10*, 2177–2191. https://doi.org/10.1029/2018MS001341

Houtekamer, P. L., & Zhang, F. (2016). Review of the ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review, 144*, 4489–4532. https://doi.org/10.1175/MWR-D-15-0440.1

Hunt, B. R., Kostelich, E. J., & Szunyogh, I. (2007). Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena, 230*(1–2), 112–126. https://doi.org/10.1016/j.physd.2006.11.008

Huntley, H. S., & Hakim, G. J. (2010). Assimilation of time-averaged observations in a quasi-geostrophic atmospheric jet model. *Climate Dynamics, 35*, 995–1009. https://doi.org/10.1007/s00382-009-0714-5

Kalnay, E. (2003). Atmospheric modeling, data assimilation and predictability. New York: Cambridge University Press.

Kalnay, E., Li, H., Miyoshi, T., Yang, S.-C., & Ballabrera-Poy, J. (2007). 4-D-Var or ensemble Kalman filter? *Tellus A: Dynamic Meteorology and Oceanography, 59*(5), 758–773. https://doi.org/10.1111/j.1600-0870.2007.00261.x

Knight, J. R., Polland, C. K., & Scaife, A. A. (2006). Climate impacts of the Atlantic Multidecadal Oscillation. *Geophysical Research Letters, 33*, L17706. https://doi.org/10.1029/2006GL026242

Kondo, K., & Miyoshi, T. (2016). Impact of removing covariance localization in an ensemble Kalman filter: Experiments with 10,240 members using an intermediate AGCM. *Monthly Weather Review, 144*, 4849–4865. https://doi.org/10.1175/MWR-D-15-0388.1

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences, 20*, 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus, 21*(3), 289–307. https://doi.org/10.3402/tellusa.v21i3.10086

Lorenz, E. N. (2006). Predictability—A problem partly solved. In T. Palmer, & R. Hagedorn (Eds.), *Predictability of weather and climate* (pp. 40–58). Cambridge University Press.

Lu, F., Liu, Z., Zhang, S., & Liu, Y. (2015a). Strongly coupled data assimilation using leading averaged coupled covariance (LACC). Part I: Simple model study. *Monthly Weather Review, 143*, 3823–3837. https://doi.org/10.1175/MWR-D-14-00322.1

Liu, J., Kalnay, E., & Cardinali, C. (2009). Analysis sensitivity calculation in an ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society, 135*, 523–533. https://doi.org/10.1002/qj.371

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences, 20*, 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus, 21*(3), 289–307. https://doi.org/10.3402/tellusa.v21i3.10086

Mathiot, P., Goosse, H., Crosta, X., Stenni, B., Braida, M., Renssen, H., et al. (2013). Using data assimilation to investigate the causes of Southern Hemisphere high latitude cooling from 10 to 8 ka BP. *Climate of the Past, 9*, 887–901. https://doi.org/10.5194/cp-9-887-2013

Matsakaris, A., Widmann, M., & Jungclaus, J. (2015). On-line and off-line data assimilation in paleoclimatology: A case study. *Climate of the Past, 11*, 81–93. https://doi.org/10.5194/cp-11-81-2015

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., et al. (2014). Decadal climate prediction: An update from the trench. *Climate Dynamics, 43*, 79–93. https://doi.org/10.1007/s00382-013-1874-7

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., et al. (2014). Decadal climate prediction: An update from the trench. *Climate Dynamics, 43*, 79–93. https://doi.org/10.1007/s00382-013-1874-7

Miyoshi, T., Kondo, K., & Imanaka, T. (2014). The 10,240-member ensemble Kalman filtering with an intermediate AGCM. *Nonlinear Processes in Geophysics, 21*, 107–126. https://doi.org/10.5194/npg-21-107-2014

Miyoshi, T., Kondo, K., & Imanaka, T. (2014). The 10,240-member ensemble Kalman filtering with an intermediate AGCM. *Nonlinear Processes in Geophysics, 21*, 107–126. https://doi.org/10.5194/npg-21-107-2014

Miyoshi, T., Kondo, K., Imanaka, T., & Yasuoka, K. (2011). Ensemble Kalman filter experiments with a primitive-equation global model. Ph.D. dissertation, University of Maryland, College Park, 197pp.

Miyoshi, T. (2011). The gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. *Monthly Weather Review, 139*, 1519–1535. https://doi.org/10.1175/2010MWR3570.1

Miyoshi, T., Kondo, K., & Imanaka, T. (2014). The 10,240-member ensemble Kalman filtering with an intermediate AGCM. *Geophysical Research Letters, 41*, 5264–5271. https://doi.org/10.1002/2014GL060863

Molteni, F. (2003). Atmospheric simulations using a GCM with simplified physical parametrizations. I: Model climatology and variability in multi-decadal experiments. *Climate Dynamics, 20*, 175–191. https://doi.org/10.1007/s00382-002-0268-2

Oke, P. R., Schiller, A., Griffin, D. A., & Brassington, G. B. (2005). Ensemble data assimilation for an eddy-resolving ocean model of the Australian region. *Quarterly Journal of the Royal Meteorological Society, 131*, 3301–3311. https://doi.org/10.1256/qj.05.95

PAGES 2k Consortium. (2013). Continental-scale temperature variability during the past two millennia. *Nature Geoscience, 6*, 339–346. https://doi.org/10.1038/ngeo1797

Pendergrass, A. G., Hakim, G. J., Battisti, D. S., & Roe, G. (2012). Coupled air-mixed layer temperature predictability for climate reconstruction. *Journal of Climate, 25*, 459–472. https://doi.org/10.1175/2011JCLI4494.1

Penny, S. G., & Hamill, T. M. (2017). Coupled Data assimilation for integrated earth system analysis and prediction. *Bulletin of the American Meteorological Society, 98*, ES169–ES172. https://doi.org/10.1175/BAMS-D-17-0036.1
Perkins, W. A., & Hakim, G. J. (2017). Reconstructing paleoclimate fields using online data assimilation with a linear inverse model. *Climate of the Past*, 13, 421–436. https://doi.org/10.5194/cp-13-421-2017

Perkins, W. A., & Hakim, G. J. (2021). Coupled atmosphere–ocean reconstruction of the last millennium using online data assimilation. *Paleoceanography and Paleoclimatology*, 36, e2020PA003959. https://doi.org/10.1029/2020PA003959

Renssen, H., Mairesse, A., Goose, H., Mathiot, P., Heiri, O., Roche, D. M., et al. (2015). Multiple causes of the Younger Dryas cold period. *Nature Geoscience*, 8, 946–949. https://doi.org/10.1038/ngeo2557

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, B., Clark, R. T., Dunstone, N., et al. (2014). Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, 41, 2514–2519. https://doi.org/10.1002/2014GL061967

Sukha, T. C., Penny, S. G., Kalnay, E., & Miyoshi, T. (2016). Assimilating atmospheric observations into the ocean using strongly coupled ensemble data assimilation. *Geophysical Research Letters*, 43, 752–759. https://doi.org/10.1002/2015GL067238

Steiger, N. J., Hakim, G. J., Steig, E. J., Battisti, D. S., & Roe, G. H. (2014). Assimilation of time-averaged pseudoproxies for climate reconstruction. *Journal of Climate*, 27, 426–441. https://doi.org/10.1175/JCLI-D-12-00693.1

Steiger, N. I., Smerdon, J. E., Cook, E. R., & Cook, B. I. (2014). Assimilation of time-averaged pseudoproxies for climate reconstruction. *Journal of Climate*, 27, 149–166. https://doi.org/10.1175/JCLI-D-13-00538.1

Tardif, R., Hakim, G. J., Perkins, W. A., Horlick, K. A., Erb, M. P., Emile-Geay, J., et al. (2019). Last Millennium Reanalysis with an expandable proxy database and seasonal proxy modeling. *Climate of the Past*, 15, 1251–1273. https://doi.org/10.5194/cp-15-1251-2019

Tardif, R., Hakim, G. J., & Snyder, C. (2014). Coupled atmosphere–ocean data assimilation experiments with a low-order climate model. *Climate Dynamics*, 43, 1631–1643. https://doi.org/10.1007/s00382-013-1989-0

Tardif, R., Hakim, G. J., & Snyder, C. (2015). Coupled atmosphere–ocean data assimilation experiments with a low-order model and CMIP5 model data. *Climate Dynamics*, 45, 1415–1427. https://doi.org/10.1007/s00382-014-2390-3

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bull. Amer. Meteorol. Soc.* 93, 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

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

Trenberth, K. E., & Caron, J. M. (2000). The southern oscillation revisited: Sea level pressures, surface temperatures and precipitation. *Journal of Climate*, 13, 4358–4365. https://doi.org/10.1175/1520-0442(2000)013<4358:TSORSL>2.0.CO;2

van der Schrier, G., & Barkmeijer, J. (2005). Bjerknes’ hypothesis on the coldness during AD 1790–1820 revisited. *Climate of the Past*, 1, 141–145. https://doi.org/10.5194/cp-1-141-2005

van Leeuwen, P. J. (2009). Particle filtering in geophysical systems. *Monthly Weather Review*, 137, 4089–4114. https://doi.org/10.1175/2009MWR2835.1

von Storch, H., Cushman, U., Gonzalez-Rouco, J. F., Jones, J. M., Voss, R., Widmann, M., & Zorita, E. (2000). Combining paleoclimatic evidence and GCMs by means of data assimilation through upscaling and nudging (DATUN). *11th American Symposium on Global Change*, 28–31. Retrieved from http://www.hvonstorch.de/klima/pdf/DATUN-2000.pdf

Watanabe, M., Suzuki, T., O’ishi, R., Komuro, Y., Watanabe, S., Emori, S., et al. (2010). Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity. *Journal of Climate*, 23, 6312–6335. https://doi.org/10.1175/2010JCLI3679.1

Whitaker, J. S., & Hamill, T. M. (2012). Evaluating methods to account for system errors in ensemble data assimilation. *Monthly Weather Review*, 140, 3078–3089. https://doi.org/10.1175/MWR-D-11-00276.1

WMO. (2017). Coupled Data assimilation for integrated earth system analysis and prediction: Goals, challenges and recommendations. WWRP 2017-3. Retrieved from https://www.wmo.int/pages/prog/arep/wrrp/new/documents/Final_WWRP_2017_3_27_July.pdf

Yoshida, T., & Kalnay, E. (2018). Correlation-cutoff method for covariance localization in strongly coupled data assimilation. *Monthly Weather Review*, 146, 2881–2899. https://doi.org/10.1175/MWR-D-17-0365.1

Zhang, P., Odinis, A. M., & Nielsen-Gammon, J. W. (2006). Mesoscale predictability of an extreme warm-season precipitation event. *Weather and Forecasting*, 21, 149–166. https://doi.org/10.1175/WAF909.1