An Online Paleoclimate Data Assimilation with a Deep Learning-based Network

Haohao Sun [1] Lili Lei [1] Zhengyu Liu [2] Liang Ning [3] Zhe-Min Tan [1]

[1]School of Atmospheric Sciences, Nanjing University, Nanjing
[2]Department of Geography, The Ohio State University, Columbus, Ohio
[3]School of Geography, Nanjing Normal University, Nanjing

Introduction

Paleoclimate data assimilation (PDA) combines information from climate simulations and proxy data to provide an optimal estimate of the past climate. Due to the tremendous computational cost required by lone-time simulations of an Earth system model, PDA has been traditionally performed in an offline scenario, in which priors are selected from existing paleoclimate simulations.

Currently, deep learning-based networks have been promptly developed, which are data-driven and computationally efficient after trained, and thus provide an alternative for online PDA. The deep learning-based network along with an integrated hybrid ensemble Kalman filter is proposed as an online PDA here. This online PDA can efficiently provide climate forecasts with predictive skills and effectively assimilate sparse proxy data, leading to well reconstructed surface air temperature.

Surrogate models

- **Linear regression model**: The linear inverse model (LIM, Penland and Sardeshmukh 1995) is a linear Markov process that represents the evolution of a dynamic system. The model state at t mapped from the model state at $t-1$ is given by

$$\mathbf{x}_t = \mathbf{M}_{LIM}\mathbf{x}_{t-1} + \sigma_{t-1} \tag{1}$$

where the linear operator matrix $\mathbf{M}_{LIM}$ is given by the regression of covariance matrices of the state vectors with time lags of 0 and 1 in the empirical orthogonal function (EOF) space.

- **Deep learning-based network**: A deep learning-based network (NET) is built to map the model state at $t-1$ to that at t, which is a convolutional decoder-encoder (fig. 1).The input and output are three-dimensional state variables, which contain augmented state variables of annual mean and six seasonal averages that are required by the proxy system models (PSMs)

$$\mathbf{x}_t = M_{NET}\left(\mathbf{x}_{t-1}\right) \tag{2}$$

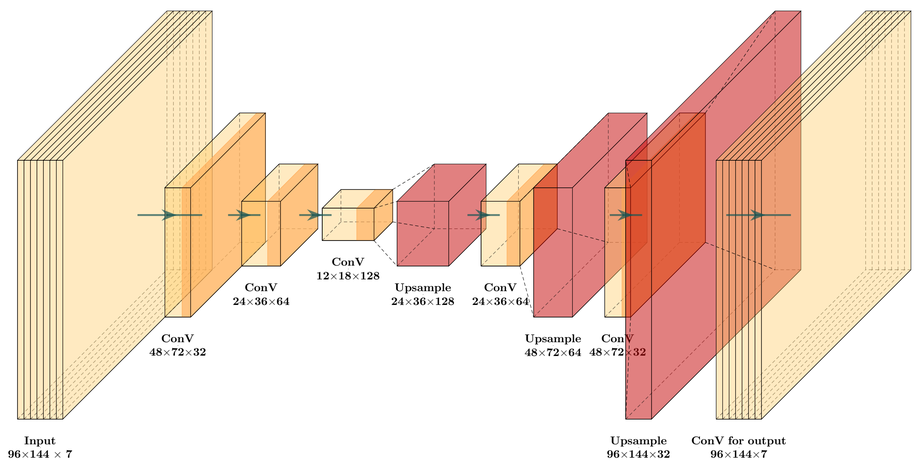

Figure 1. Architecture of the convolutional decoder-encoder network. The encoder comprises one input layer, three convolutional layers (ConV), followed by three upsample-convolutional layers for the decoder. The activation layer is denoted by a darker orange shading.

Data assimilation methods

- **Offline data assimilation**: The offline EnKF [Hakim et al. 2016] is used as a benchmark. Ensemble priors $\mathbf{x}^f$ are randomly sampled from a climatological simulation. The ensemble-square root filter (EnSRF), as a deterministic filter, updates the ensemble mean and ensemble perturbations separately.

$$\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^f + \mathbf{P}^f\mathbf{H}^{\mathrm{T}}\left[\mathbf{H}\mathbf{P}^f\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right]^{-1}\left[\mathbf{y} - H\left(\bar{\mathbf{x}}^f\right)\right] \tag{3}$$

$$\mathbf{x}'^a_i = \mathbf{x}'^f_i - \mathbf{P}^f\mathbf{H}^{\mathrm{T}}\left[\left(\sqrt{\mathbf{H}\mathbf{P}^f\mathbf{H}^{\mathrm{T}} + \mathbf{R}}\right)^{-1}\right]^{\mathrm{T}}\left[\sqrt{\mathbf{H}\mathbf{P}^f\mathbf{H}^{\mathrm{T}} + \mathbf{R}} + \sqrt{\mathbf{R}}\right]^{-1}\left(\mathbf{H}\mathbf{x}'^f_i\right) \tag{4}$$

- **Online data assimilation**: Ensemble priors $\mathbf{x}^f_{cyc}$ of online DA are short-term forecasts based on the surrogate models. Ensemble mean and ensemble perturbations are updated separately by the hybrid background error covariances through integrated hybrid EnKF (IHEnKF, Lei et al. 2021). The hybrid background error covariances $\mathbf{P}^f_{hyb}$ are a combination of the flow-dependent background error covariances $\mathbf{P}^f_{cyc}$ and the static background error covariances $\mathbf{P}^f_{climo}$, with weights α and $1-\alpha$ respectively.

$$\bar{\mathbf{x}}^a_{cyc} = \bar{\mathbf{x}}^f_{cyc} + \mathbf{P}^f_{hyb}\mathbf{H}^{\mathrm{T}}\left[\mathbf{H}\mathbf{P}^f_{hyb}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right]^{-1}\left[\mathbf{y} - H\left(\bar{\mathbf{x}}^f_{cyc}\right)\right] \tag{5}$$

$$\mathbf{x}'^a_{cyc,i} = \mathbf{x}'^f_{cyc,i} - \mathbf{P}^f_{hyb}\mathbf{H}^{\mathrm{T}}\left[\left(\sqrt{\mathbf{H}\mathbf{P}^f_{hyb}\mathbf{H}^{\mathrm{T}} + \mathbf{R}}\right)^{-1}\right]^{\mathrm{T}}\left[\sqrt{\mathbf{H}\mathbf{P}^f_{hyb}\mathbf{H}^{\mathrm{T}} + \mathbf{R}} + \sqrt{\mathbf{R}}\right]^{-1}\left(\mathbf{H}\mathbf{x}'^f_{cyc,i}\right) \tag{6}$$

Since the surrogate model tends to lost ensemble variances along with lead times, a blended prior is adopted [Perkins and Hakim 2017]. The selected climatological $\mathbf{x}^f_{\mathrm{climo}}$ or analog priors $\mathbf{x}^f_{\mathrm{analog}}$ can be added to the ensemble priors $\mathbf{x}^f_{cyc}$ with coefficient β:

$$\mathbf{x}^f_{cyc,i} = \beta M\left(\mathbf{x}^a_{cyc,i}\right) + (1-\beta)\,\mathbf{x}^f_i \tag{7}$$

Experimental design

- **Proxy and simulation data**: The proxy records from the PAGES 2k Consortium [PAGES2k Consortium et al. 2017] that are temperature sensitive are assimilated, with temporal coverage of the Common Era (0-2000CE). The maximum number of proxies in 1945 CE is 554 (fig. 2).
- **Pseudoproxy and real proxy experiments**: Pseudoproxy data is generated by adding random error from a normal distribution to the true observed value; Real proxies from PAGES 2k are further utilized; Six different proxy networks with 50, 150, 250, 350, 450, and the full proxies are tested.
- **Evaluation metrics**: Root-mean square error (RMSE) and coefficient of efficiency (CE) are computed.

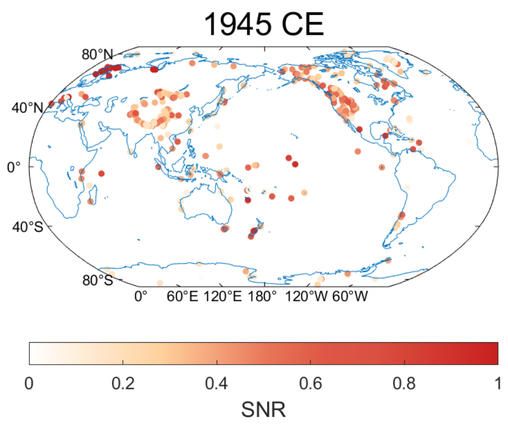

Figure 2. Spatial distributions of proxy records in 1945 CE with colors indicating the signal-to-noise ratio

Predictive skills of surrogate models

- Compared to the offline sampling, LIM has improved predictive skills within five years, and the shorter the lead time the better the predictive skill.
- Compared to the LIM, the NET generally further improves the predictive skills.

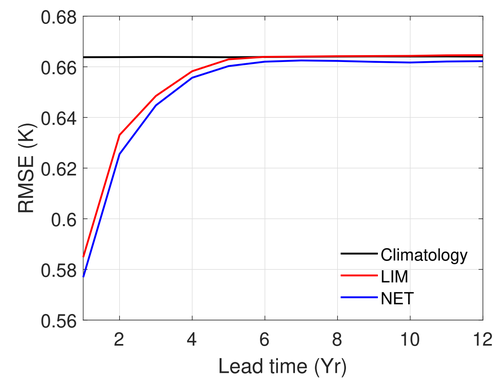

Figure 3. Spatially and temporally averaged RMSEs for climatology, LIM, and NET at various forecast lead times (years)

Results from pseudoproxy experiments

- Exp_LIM and Exp_NET have larger CEs than Exp_OFF, which indicates improved reconstruction skills from the online DA with surrogate models than the offline DA.
- Compared to Exp_LIM, Exp_NET obtains larger CEs of priors and posteriors, the improved predictive skills contributed by the NET than the LIM are beneficial for online DA.
- The differences between Exp_LIM and Exp_NET are enlarged with sparse proxy data.

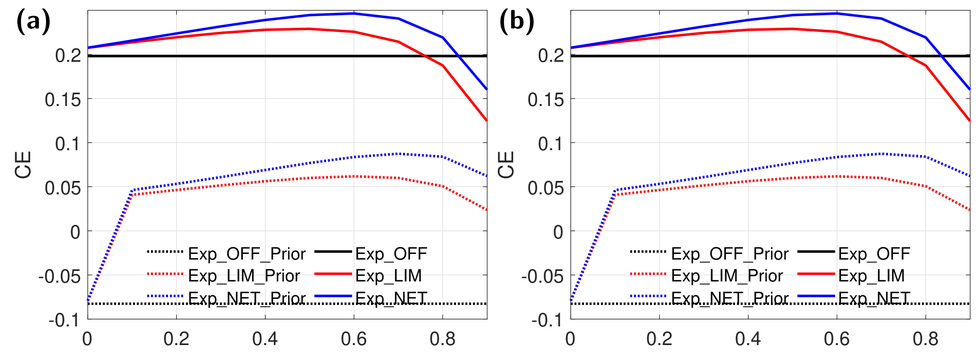

Figure 4. Area-weighted global averaged CEs for Exp_OFF, Exp_LIM, and Exp_NET with varying hybrid weights / blending weights, assimilating a proxy network with (a) 150 pseudoproxies and (b) full pseudoproxies.

Results from real proxy experiments

- When a full set of proxy data is used, offline DA and online DA with either LIM or NET have similar GMT reconstructions, which are mainly forced by the proxy information.
- When a limited set of proxies is used, Exp_OFF has reduced GMT variability than Exp_LIM and Exp_NETANA.
- Compared to Exp_LIM with linear predictive skills obtained by the LIM, Exp_NETANA benefits from retaining nonlinear dynamics provided by the NET and further incorporating analog climatological ensembles, which leads to improved GMT reconstruction even with limited proxies.

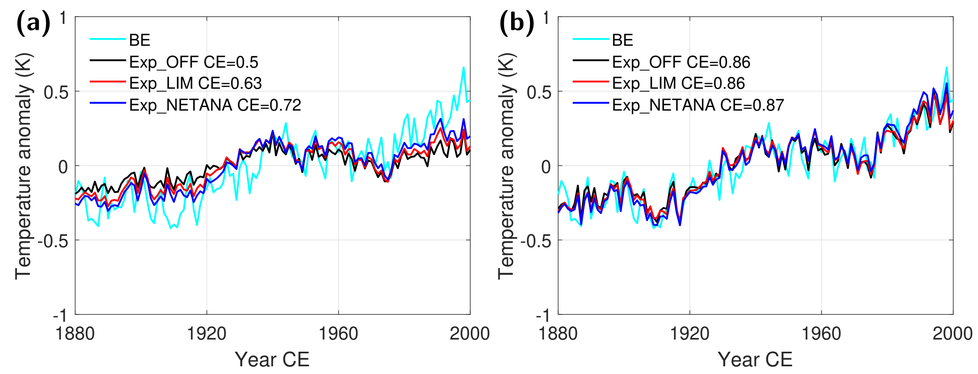

Figure 5. The reconstruction of GMTs over the instrumental period for Exp_OFF, Exp_LIM, and Exp_NETANA given a hybrid weight / blending weight of 0.7, assimilating a proxy network with (a) 150 proxies and (b) full proxies.

- The GMT reconstruction provided by the online PDA has smaller spread but more consistent error statistics than those provided by the online PDA with LIM and offline PDA, during the early period with limited proxy data.

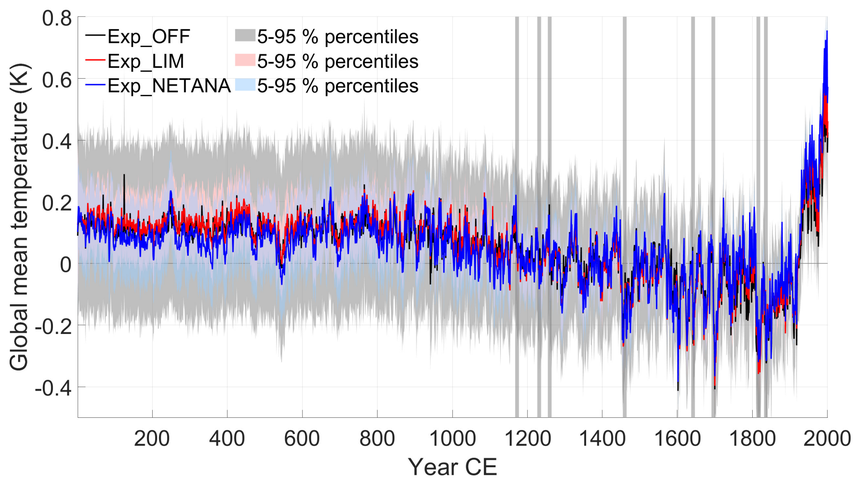

Figure 6. The reconstruction of GMTs over the Common Era for Exp_OFF, Exp_LIM, and Exp_NETANA, given a hybrid weight / blending weight of 0.7, assimilating a proxy network with full proxies. The vertical grey lines denote the volcanic eruption years.