Data Assimilation with δ^{18}O from isotope-enabled GCMs and terrestial paleoclimatic archives for the last millennium

Master thesis project

Mathurin Choblet

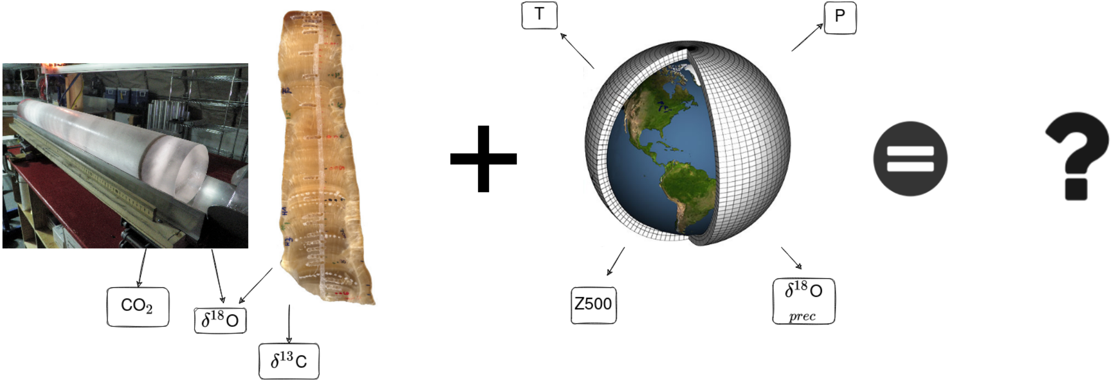
mchoblet@iup.uni-heidelberg.de

February 9, 2022

Table of contents

1. Motivation
2. From the Kalman Filter towards Paleoclimate DA
3. The Last Millennium Reanalysis framework (LMR)
4. Data
5. Research questions
6. Project plan

Data Assimilation for Paleoclimate-Research



CO_2

$\delta^{18}O$

$\delta^{13}C$

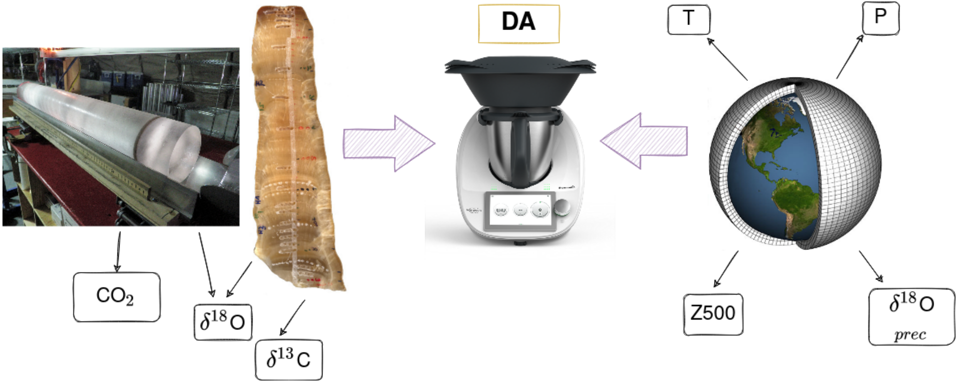
T

P

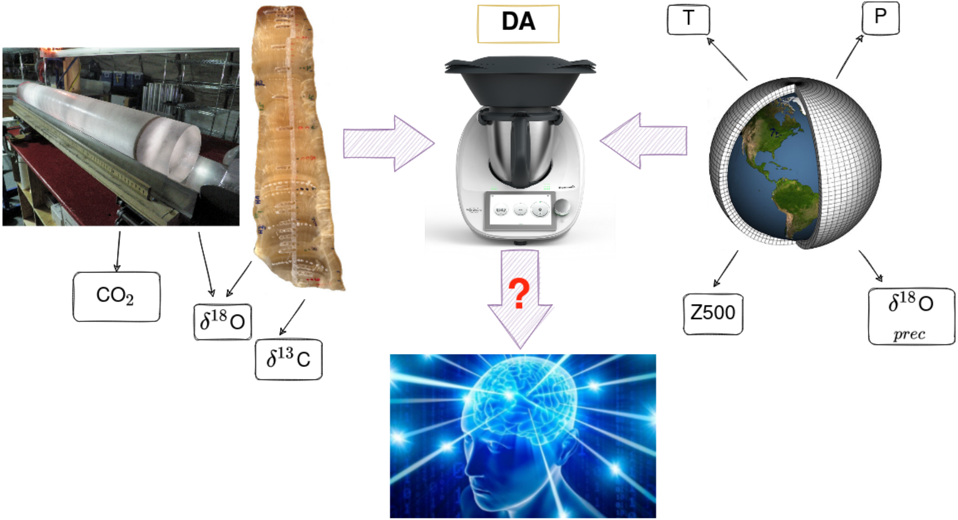
Z500

$\delta^{18}O_{prec}$

Data Assimilation for Paleoclimate-Research



Data Assimilation for Paleoclimate-Research



CO_2

$\delta^{18}O$

$\delta^{13}C$

DA

?

T

P

Z500

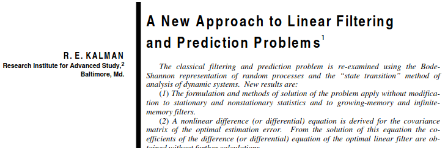
$\delta^{18}O_{prec}$

What is the Kalman Filter (KF)?

The answer to the question:

Suppose you have a theoretical **model** and **observations**:

What is the best way of combining both in order to minimize the mean squared error?



R. E. KALMAN
Research Institute for Advanced Study,[2]
Baltimore, Md.

A New Approach to Linear Filtering and Prediction Problems[1]

The classical filtering and prediction problem is re-examined using the Bode-Shannon representation of random processes and the "state transition" method of analysis of dynamic systems. New results are:

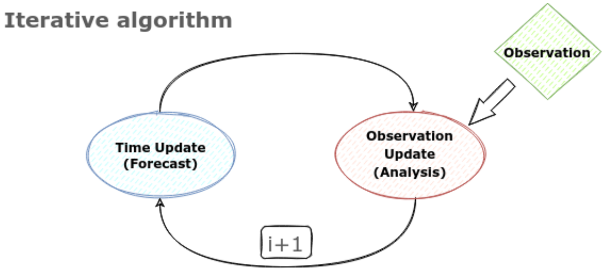
(1) The formulation and methods of solution of the problem apply without modification to stationary and nonstationary statistics and to growing-memory and infinite-memory filters.

(2) A nonlinear difference (or differential) equation is derived for the covariance matrix of the optimal estimation error. From the solution of this equation the coefficients of the difference (or differential) equation of the optimal linear filter are obtained without further calculations.

Figure 1: Original publication [Kalman, 1960]

→ The KF provides an optimal weighted mean which also minimizes the uncertainty (error variance).

Iterative Kalman Filter scheme

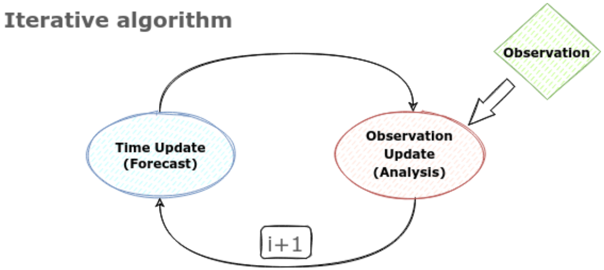


Iterative Kalman Filter scheme

Forecast equations

$$x_i^f = Fx_{i-1} \tag{1}$$

$$B_i^f = FB_{i-1}F^T + Q \tag{2}$$



Definitions

x	state vector	H	obs. operator
z	observation	B	error covariance
F	linear model	Q, R	model/obs. err.
K	Kalman gain		

Iterative Kalman Filter scheme

Forecast equations

$$x_i^f = F x_{i-1} \tag{1}$$

$$B_i^f = F B_{i-1} F^T + Q \tag{2}$$

Analysis equations

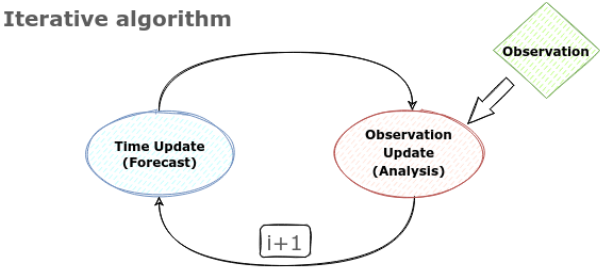
$$x_i^a = x_i^f + K_i (z_i - H x_i^f) \tag{3}$$

$$K_i = B_i^f H^T (H B_i^f H^T + R)^{-1} \tag{4}$$

$$B_i^a = (I - K_i H) B_i^f \tag{5}$$

Definitions

x	state vector
z	observation
F	linear model
K	Kalman gain
H	obs. operator
B	error covariance
Q, R	model/obs. err.



Iterative Kalman Filter scheme

Forecast equations

$$x_i^f = F x_{i-1} \tag{1}$$

$$B_i^f = F B_{i-1} F^T + Q \tag{2}$$

Analysis equations

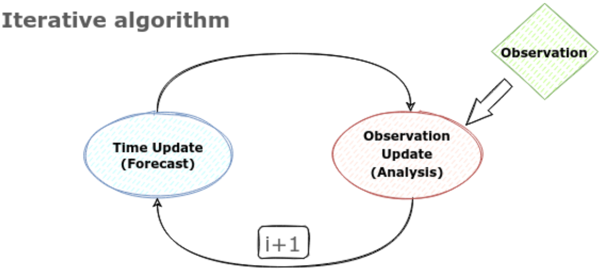
$$x_i^a = x_i^f + K_i(z_i - H x_i^f) \tag{3}$$

$$K_i = B_i^f H^T (H B_i^f H^T + R)^{-1} \tag{4}$$

$$B_i^a = (I - K_i H) B_i^f \tag{5}$$

Definitions

x	state vector
z	observation
F	linear model
K	Kalman gain
H	obs. operator
B	error covariance
Q, R	model/obs. err.



Requirements

- Gaussian errors
- Linear forward model

The Ensemble Kalman Filter (EnKF)
A Monte Carlo based implementation (Evensen 1994)

- Kalman Filter not suited for nonlinear models
  (What's the covariance?)

The Ensemble Kalman Filter (EnKF)
A Monte Carlo based implementation (Evensen 1994)

- Kalman Filter not suited for nonlinear models
(What's the covariance?)
- State vector x is replaced by matrix X for an ensemble of N models

$$\mathrm{x} = \begin{bmatrix} x_1 \\ \ldots \\ x_m \end{bmatrix} \rightarrow \mathrm{X} = \begin{bmatrix} x_{11} & \ldots & x_{1N} \\ x_{i1} & \ldots & x_{iN} \\ x_{m1} & \ldots & x_{mN} \end{bmatrix} \qquad (6)$$

- Approximate model error covariance through ensemble
$B_i^f = \frac{1}{N-1}(X - \bar{X})(X - \bar{X})^T$

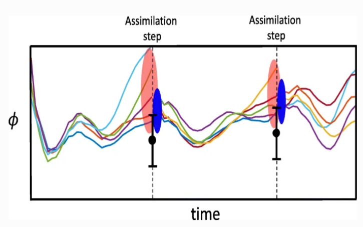


Figure 2: Principle of EnKF [Labahn et al., 2020]

The Ensemble Kalman Filter (EnKF)
A Monte Carlo based implementation (Evensen 1994)

- Kalman Filter not suited for nonlinear models (What's the covariance?)
- State vector x is replaced by matrix X for an ensemble of N models

$$\mathrm{x} = \begin{bmatrix} x_1 \\ \ldots \\ x_m \end{bmatrix} \rightarrow \mathrm{X} = \begin{bmatrix} x_{11} & \ldots & x_{1N} \\ x_{i1} & \ldots & x_{iN} \\ x_{m1} & \ldots & x_{mN} \end{bmatrix} \tag{6}$$

- Approximate model error covariance through ensemble $B_i^f = \frac{1}{N-1}(X - \bar{X})(X - \bar{X})^T$
- Kalman equations stay the same

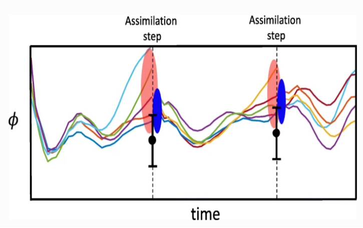


Figure 2: Principle of EnKF [Labahn et al., 2020]

EnKF applied to the Lorenz 63 model

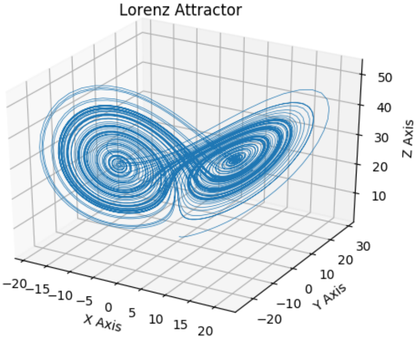
L63 system

$$\dot{x} = -\sigma(x - y) \tag{7}$$

$$\dot{y} = x(\rho - z) - y \tag{8}$$

$$\dot{z} = xy - bz \tag{9}$$

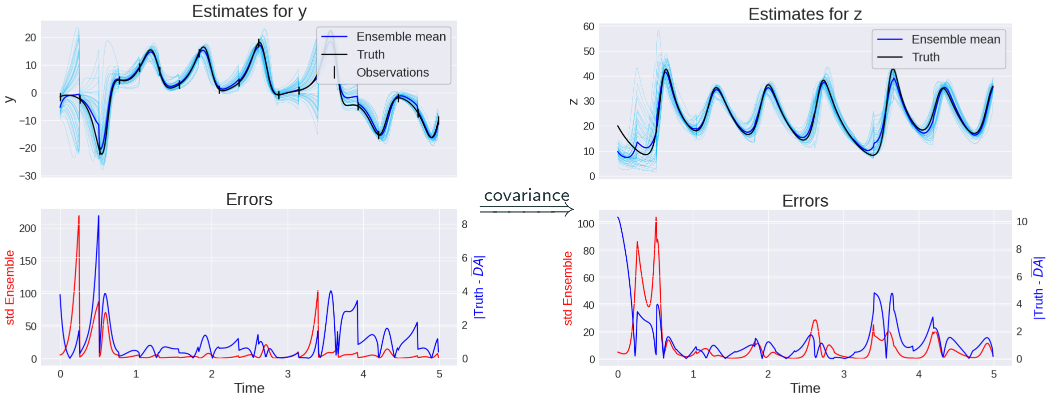
e.g. $\sigma = 10$, $\rho = 28$ and $b = 8/3$



Measuring only one of the three variables, can we estimate the whole trajectory?

EnKF applied to the Lorenz 63 model

Reconstruction using measurements in y and wrong initial conditions



- EnKF could also be used for parameter estimation

Sequential Proxy Assimilation

Compute K for each proxy location

$$\begin{aligned} K &= \mathrm{BH}^{\mathrm{T}}(\mathrm{HBH}^{\mathrm{T}} + R)^{-1} \tag{10} \\ &= \mathrm{cov}(\mathrm{X}, \mathrm{HX})[\mathrm{cov}(\mathrm{HX}, \mathrm{HX}) + R]^{-1} \tag{11} \\ &= \frac{\mathrm{cov}(\mathrm{X}, \mathrm{HX})}{\mathrm{var}(\mathrm{HX}) + R} \tag{12} \end{aligned}$$

with

- X State vector/matrix from model
- H Observation Operator
- Hx model prior at proxy location
- R measurement error

Sequential Proxy Assimilation

Compute K for each proxy location

$$K = \mathrm{BH}^{\mathrm{T}}(\mathrm{HBH}^{\mathrm{T}} + R)^{-1} \tag{10}$$

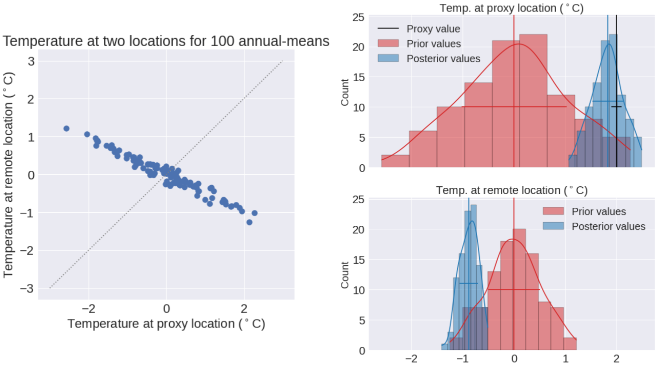
$$= \mathrm{cov}(\mathrm{X}, \mathrm{HX})[\mathrm{cov}(\mathrm{HX}, \mathrm{HX}) + R]^{-1} \tag{11}$$

$$= \frac{\mathrm{cov}(\mathrm{X}, \mathrm{HX})}{\mathrm{var}(\mathrm{HX}) + R} \tag{12}$$

with

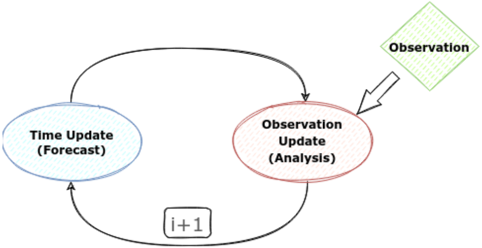
- X State vector/matrix from model
- H Observation Operator
- Hx model prior at proxy location
- R measurement error

Example: T at two locations, only T_1 measured.



Offline Data-Assimilation

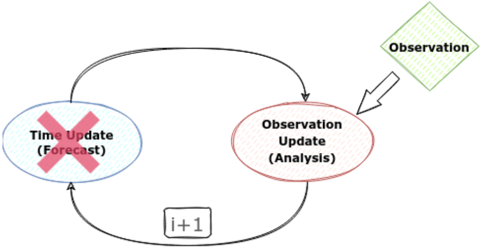
- Proxy record represents time-averaged climate state variable
- Cycling approach computationally costly for GCMs
- Predictability issue for long time-scales: Models climatology may be as informative as model forecast



Offline Data-Assimilation

- Proxy record represents time-averaged climate state variable
- Cycling approach computationally costly for GCMs
- Predictability issue for long time-scales: Models climatology may be as informative as model forecast

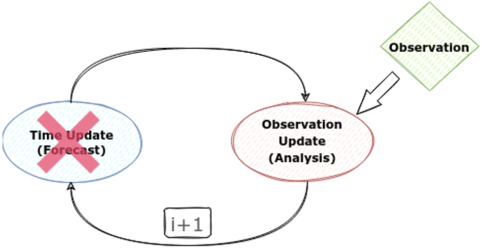
→ Omit the forecast step. Do not reinitialize the model.



Offline Data-Assimilation

- Proxy record represents time-averaged climate state variable
- Cycling approach computationally costly for GCMs
- Predictability issue for long time-scales: Models climatology may be as informative as model forecast

→ Omit the forecast step. Do not reinitialize the model.



- **Transient** Offline DA: Use time series of the model
- **Stationary** Offline DA: Use all modeled years as one prior. Give up temporal consistency, but: huge prior ensemble to sample from

The Last Millennium Reanalysis framework [Hakim et al., 2016]

LMR characterstics

- Proxy data: Pages2k (Annual)
- Model: MPI-ESM, CESM
- Time: Last Millenium
- Reconstructed Variables: T, Z500
- Method: Stationary Offline DA, PSMs

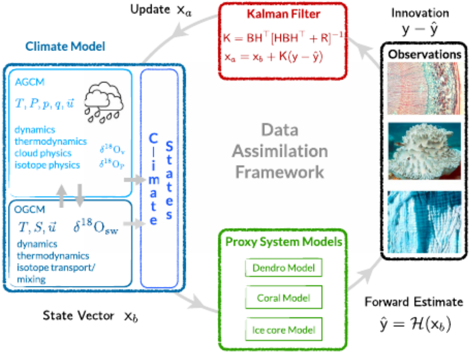


Figure 3: Kalman Filter Analysis cycle in LMR Project [Hakim et al., 2016]

The Last Millennium Reanalysis framework [Hakim et al., 2016]

LMR characterstics

- Proxy data: Pages2k (Annual)
- Model: MPI-ESM, CESM
- Time: Last Millenium
- Reconstructed Variables: T, Z500
- Method: Stationary Offline DA, PSMs

Possible enhancements

- Not only annually resolved proxies

→ Speleothems!

- Isotope enabled models

→ δ^{18}O assimilation

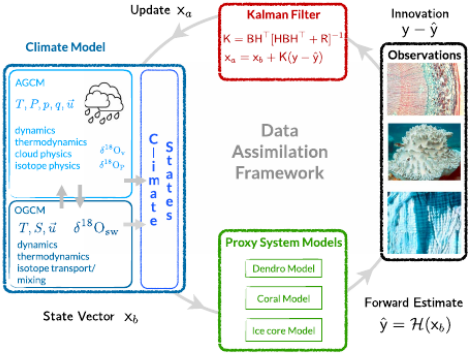


Figure 3: Kalman Filter Analysis cycle in LMR Project [Hakim et al., 2016]

Data for my project

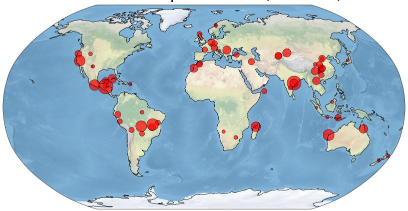
Proxies

- Speleothems (SISAL v2)
- Icecores (Iso2k)

Models [Bühler et al., 2021]

- ECHAM5-wiso
- GISS
- iCESM
- iHadCM3
- isoGSM

Locations of speleothems (SISAL 1k)



Increasing dotsize represents datapoints per location (min=39, max=1738)

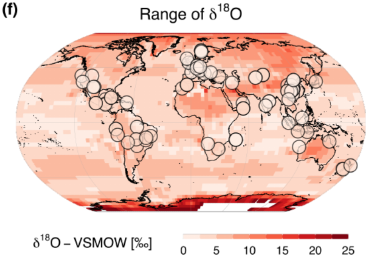


Figure 4: Multi-model range[Bühler et al., 2021]

1. How to deal with the irregular time series?

Proxies used in LMR (2019v):

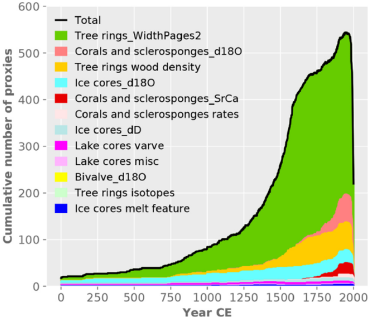
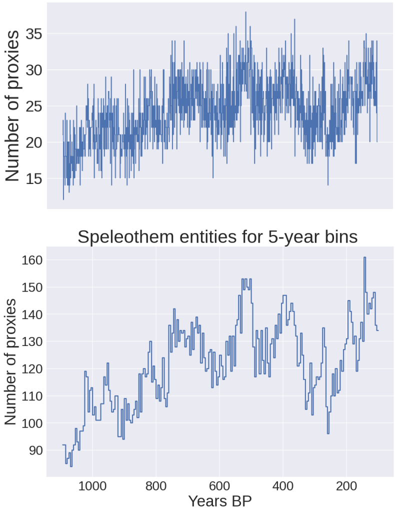


Figure 5: [Tardif et al., 2019]

- Trade-Off: Number of proxies - Time resolution
- Downsampling of prior ensemble

Speleothem availability:



2. How to deal with the spatial sparsity of proxies?

Proxies used in LMR (2019v):

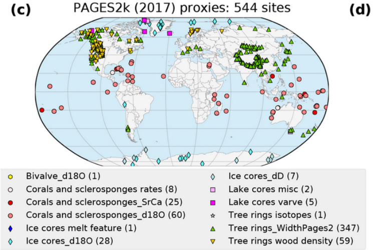
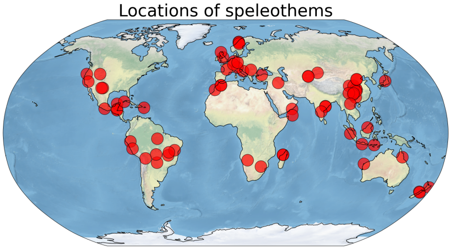

(c)
PAGES2k (2017) proxies: 544 sites
(d)
○ Bivalve_d18O (1)
○ Corals and sclerosponges rates (8)
● Corals and sclerosponges_SrCa (25)
● Corals and sclerosponges_d18O (60)
◆ Ice cores melt feature (1)
◇ Ice cores_d18O (28)
◇ Ice cores_dD (7)
□ Lake cores misc (2)
■ Lake cores varve (5)
☆ Tree rings isotopes (1)
△ Tree rings_WidthPages2 (347)
▽ Tree rings wood density (59)

Figure 6: [Tardif et al., 2019]

Speleothem availability:



- Sensitivity to observations
- Number of proxies vs location

→ Pseudoproxy experiments

3. What is the climatic importance of low-/high-latitude proxies?

How representative are speleothemes and icecores?

How do reconstructions compare for different variables (T, Z500, δ^{18}O)?

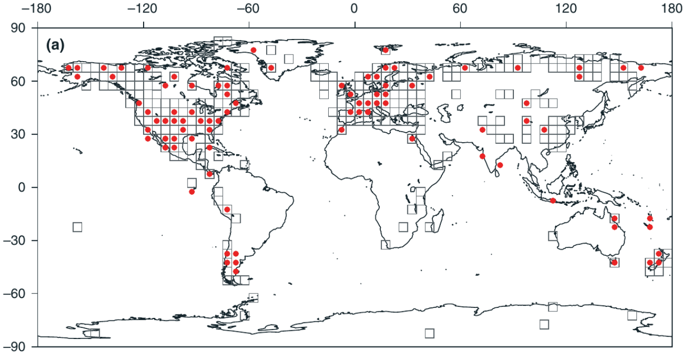


Figure 7: M. Mann Pseudoproxy Networks (grey boxes/red dots), from [Smerdon, 2012]

4. Variance of recorded and simulated δ^{18}O

Model-Data mismatch

- On decadal timescales more variabilty in Speleothemes than in models
- Literature: mismatch increases on longer timescales [Laepple and Huybers, 2014]

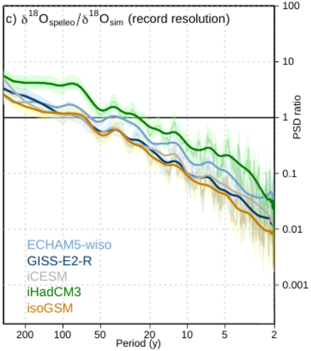


Figure 8: from Bühler 2021

How will this change for data assimilated δ^{18}O-fields?

5. Can we join the prior ensemble of different models?

- Perform reconstructions with single model/multi-model prior

Parsons et al. (2021):

We find that reconstructions derived from multi-model ensembles produce lower error than reconstructions from single-model ensembles when reconstructing independent model and instrumental data. Specifically, we find the largest decreases in error over regions far from proxy locations.

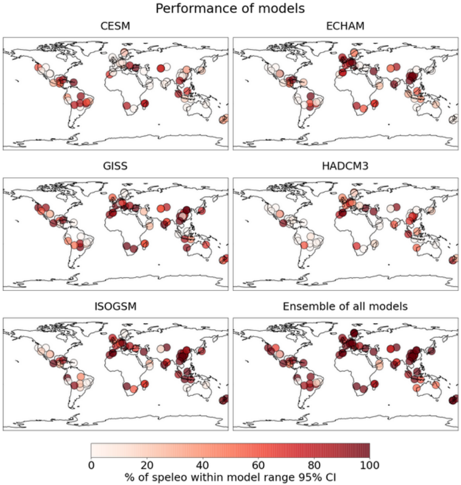


Figure 9: Model vs proxy range for our data

Project Plan

1. Literature research ✓✓
2. Basic preparation ✓✓
   - understand math of the method ✓
   - EnKF implementation for L63 ✓✓
   - proxy/model data crunching ✓✓
3. Review LMR code ✓
   - find/write working code ϰ
   - test code ϰ
   - PSM implementation ϰ
4. PPE experiments ϰ
5. Real proxies ϰ
6. Analysis/write up ϰ

Questions and discussion

Backup slides

Examples of Paleoclimate DA (1)

DA from LMR included in Pages2k Consortium Publication:

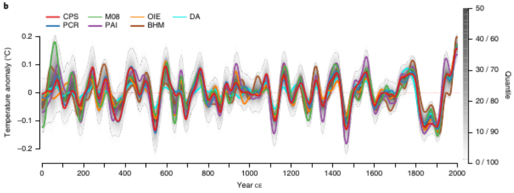


Figure 10: [Neukom et al., 2019] [Hakim et al., 2016]

[Steiger et al., 2014] did the prework for LMR and compared DA climate field reconstruction to PCA method.

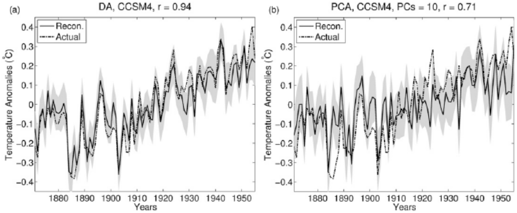
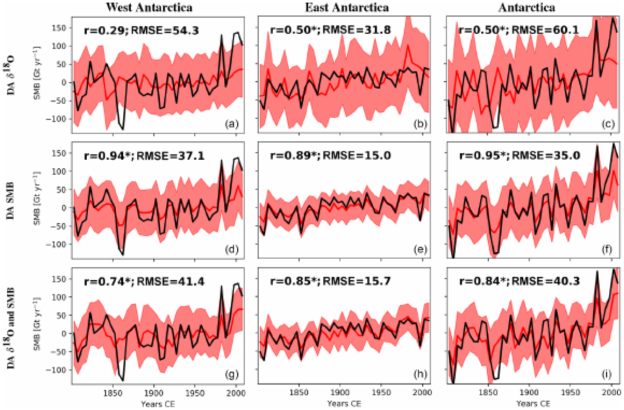


FIG. 5. Global-mean temperature anomaly reconstructions using (a) DA and (b) PCA techniques with CCSM4 over the same calibration and reconstruction periods as in Figs. 3 and 4 (calibration period: 1956–2005; reconstruction period: 1871–1955). Gray shading is one std dev of the 30 reconstructions.

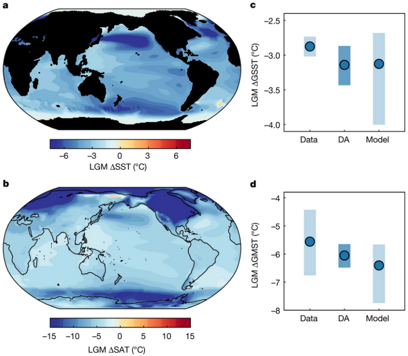
Examples of Paleoclimate DA (2)

Data Assimilation of Surface mass balance and δ^{18}O for the AIS, [Dalaiden et al., 2021]



Examples of Paleoclimate DA (3)

Application of LMR framework for the LGM [Tierney et al., 2020], [Osman et al., 2021]



a, LGM–late Holocene ΔSST. **b**, LGM–late Holocene ΔSAT. **c**, LGM ΔGSST. **d**, LGM ΔGMST. In **c** and **d**, dots represent the median values and bars the 95% CI for values derived from the data, the data assimilation (DA) and the model prior (Model).

Figure 11: Tierney 2021, LGM reconstruction

- One Model only (iCESM)
- Marine proxies only

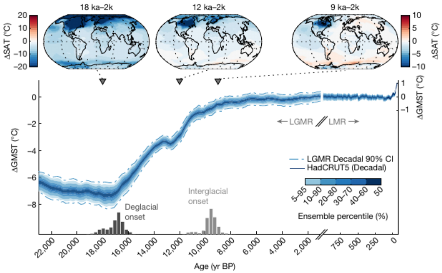
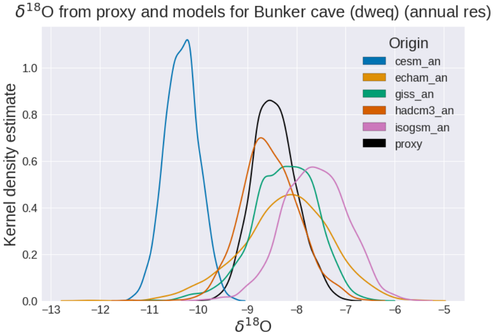
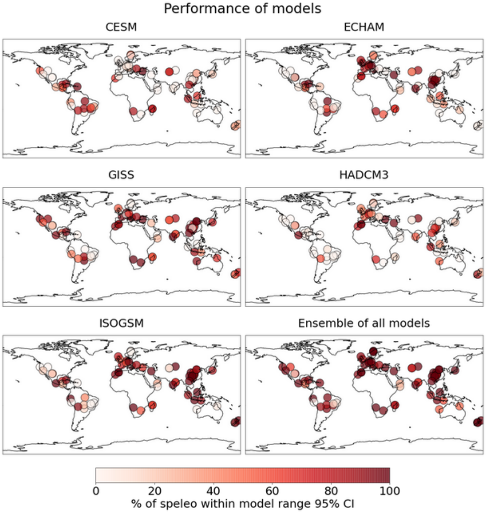


Figure 12: Osman 2021, Deglacation reconstruction

Model-Data ensemble spread comparison



Infiltration weighted. Drip water conversion uses temperature from all simulations



→ Apply bias-correction to model prior?

Verification of reconstructions

Problem

To what shall we compare the reconstructed δ^{18}O-fields to?

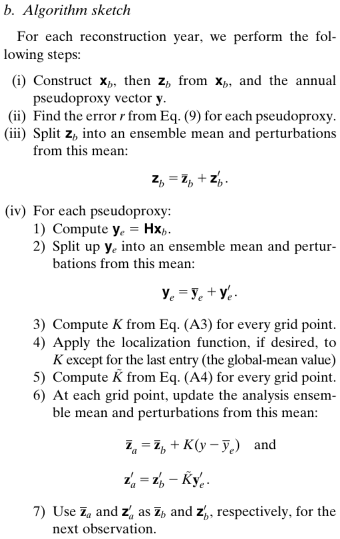
1. Compare T for instrumental period (indirect comparison)
2. Assume one model as truth, synthesize pseudoproxies (direct δ^{18}O comparison)

Metrics

- Correlation
- Coefficient of Efficiency
  (cf true time series to estimated time series) $CE = 1 - \frac{\sum(x_i - \hat{x}_i)^2}{\sum(x_i - \bar{x})^2}$
- Rank histograms (ensemble bias)
- Withholding proxy-data (cross validation)

Ensemble Square Root Kalman Filter (EnSRF)

From [Steiger et al., 2014]:



b. Algorithm sketch

For each reconstruction year, we perform the following steps:

(i) Construct $\mathbf{x}_b$, then $\mathbf{z}_b$ from $\mathbf{x}_b$, and the annual pseudoproxy vector $\mathbf{y}$.

(ii) Find the error r from Eq. (9) for each pseudoproxy.

(iii) Split $\mathbf{z}_b$ into an ensemble mean and perturbations from this mean:

$$\mathbf{z}_b = \bar{\mathbf{z}}_b + \mathbf{z}'_b .$$

(iv) For each pseudoproxy:

1) Compute $\mathbf{y}_e = \mathbf{H}\mathbf{x}_b$.
2) Split up $\mathbf{y}_e$ into an ensemble mean and perturbations from this mean:

$$\mathbf{y}_e = \bar{\mathbf{y}}_e + \mathbf{y}'_e .$$

3) Compute K from Eq. (A3) for every grid point.
4) Apply the localization function, if desired, to K except for the last entry (the global-mean value)
5) Compute $\tilde{K}$ from Eq. (A4) for every grid point.
6) At each grid point, update the analysis ensemble mean and perturbations from this mean:

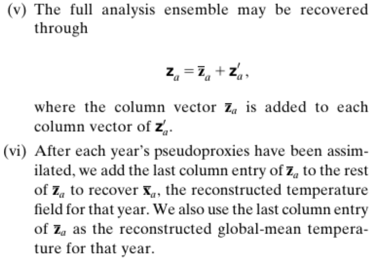
$$\bar{\mathbf{z}}_a = \bar{\mathbf{z}}_b + K(y - \bar{y}_e) \quad \text{and}$$

$$\mathbf{z}'_a = \mathbf{z}'_b - \tilde{K}\mathbf{y}'_e .$$

7) Use $\bar{\mathbf{z}}_a$ and $\mathbf{z}'_a$ as $\bar{\mathbf{z}}_b$ and $\mathbf{z}'_b$, respectively, for the next observation.

Motivation

- Ensemble Kalman Filter requires perturbed observation for ensemble variance to be correct
- EnSRF does not require this
- The main idea is to treat ensemble mean and perturbations separately (Formulas with different Kalman gains)



(v) The full analysis ensemble may be recovered through

$$\mathbf{z}_a = \bar{\mathbf{z}}_a + \mathbf{z}'_a ,$$

where the column vector $\bar{\mathbf{z}}_a$ is added to each column vector of $\mathbf{z}'_a$.

(vi) After each year's pseudoproxies have been assimilated, we add the last column entry of $\bar{\mathbf{z}}_a$ to the rest of $\bar{\mathbf{z}}_a$ to recover $\bar{\mathbf{x}}_a$, the reconstructed temperature field for that year. We also use the last column entry of $\bar{\mathbf{z}}_a$ as the reconstructed global-mean temperature for that year.

Alternative Method: The particle filter (sequential Monte Carlo)

Principle

1. Use ensemble of models (particles)
2. At each assimilation step compute RMSE between climate field of models and proxy
3. Select best fitting particle(s)
4. Resample and compute weighted mean

- Simpler concept
- More flexible, no need for normal distribution of errors
- Degeneracy problem: One particle gets all the weight

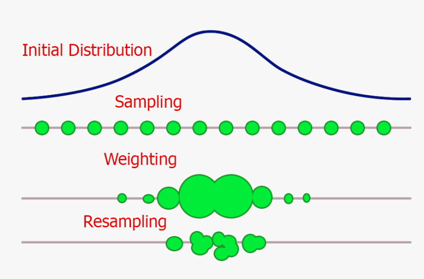


Figure 13: Particle Filter steps (without posterior distribution), from miro.medium.com

References i

J. C. Bühler, C. Roesch, M. Kirschner, L. Sime, M. D. Holloway, and K. Rehfeld. Comparison of the oxygen isotope signatures in speleothem records and ihadcm3 model simulations for the last millennium. *Climate of the Past*, 17(3):985–1004, 2021.

Q. Dalaiden, H. Goosse, J. Rezsöhazy, and E. R. Thomas. Reconstructing atmospheric circulation and sea-ice extent in the west antarctic over the past 200 years using data assimilation. *Climate Dynamics*, 57(11):3479–3503, 2021.

G. J. Hakim, J. Emile-Geay, E. J. Steig, D. Noone, D. M. Anderson, R. Tardif, N. Steiger, and W. A. Perkins. The last millennium climate reanalysis project: Framework and first results. *Journal of Geophysical Research: Atmospheres*, 121 (12):6745–6764, 2016.

References ii

R. E. Kalman. A new approach to linear filtering and prediction problems. 1960.

J. W. Labahn, H. Wu, S. R. Harris, B. Coriton, J. H. Frank, and M. Ihme. Ensemble kalman filter for assimilating experimental data into large-eddy simulations of turbulent flows. *Flow, Turbulence and Combustion*, 104(4):861–893, 2020.

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

R. Neukom, L. A. Barboza, M. P. Erb, F. Shi, J. Emile-Geay, M. N. Evans, J. Franke, D. S. Kaufman, L. Lücke, K. Rehfeld, et al. Consistent multi-decadal variability in global temperature reconstructions and simulations over the common era. *Nature geoscience*, 12(8):643, 2019.

References iii

M. B. Osman, J. E. Tierney, J. Zhu, R. Tardif, G. J. Hakim, J. King, and C. J. Poulsen. Globally resolved surface temperatures since the last glacial maximum. *Nature*, 599(7884):239–244, 2021.

J. E. Smerdon. Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments. *Wiley Interdisciplinary Reviews: Climate Change*, 3(1): 63–77, 2012.

N. J. Steiger, G. J. Hakim, E. J. Steig, D. S. Battisti, and G. H. Roe. Assimilation of time-averaged pseudoproxies for climate reconstruction. *Journal of Climate*, 27(1): 426–441, 2014.

References iv

R. Tardif, G. J. Hakim, W. A. Perkins, K. A. Horlick, M. P. Erb, J. Emile-Geay, D. M. Anderson, E. J. Steig, and D. Noone. Last millennium reanalysis with an expanded proxy database and seasonal proxy modeling. *Climate of the Past*, 15(4):1251–1273, 2019.

J. E. Tierney, J. Zhu, J. King, S. B. Malevich, G. J. Hakim, and C. J. Poulsen. Glacial cooling and climate sensitivity revisited. *Nature*, 584(7822):569–573, 2020.