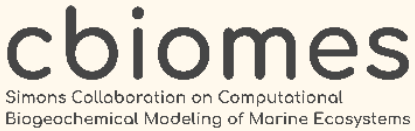
A four-dimensional ensemble optimal interpolation approach for adjoint-free biogeochemical data assimilation

J. Paul Mattern Christopher A. Edwards

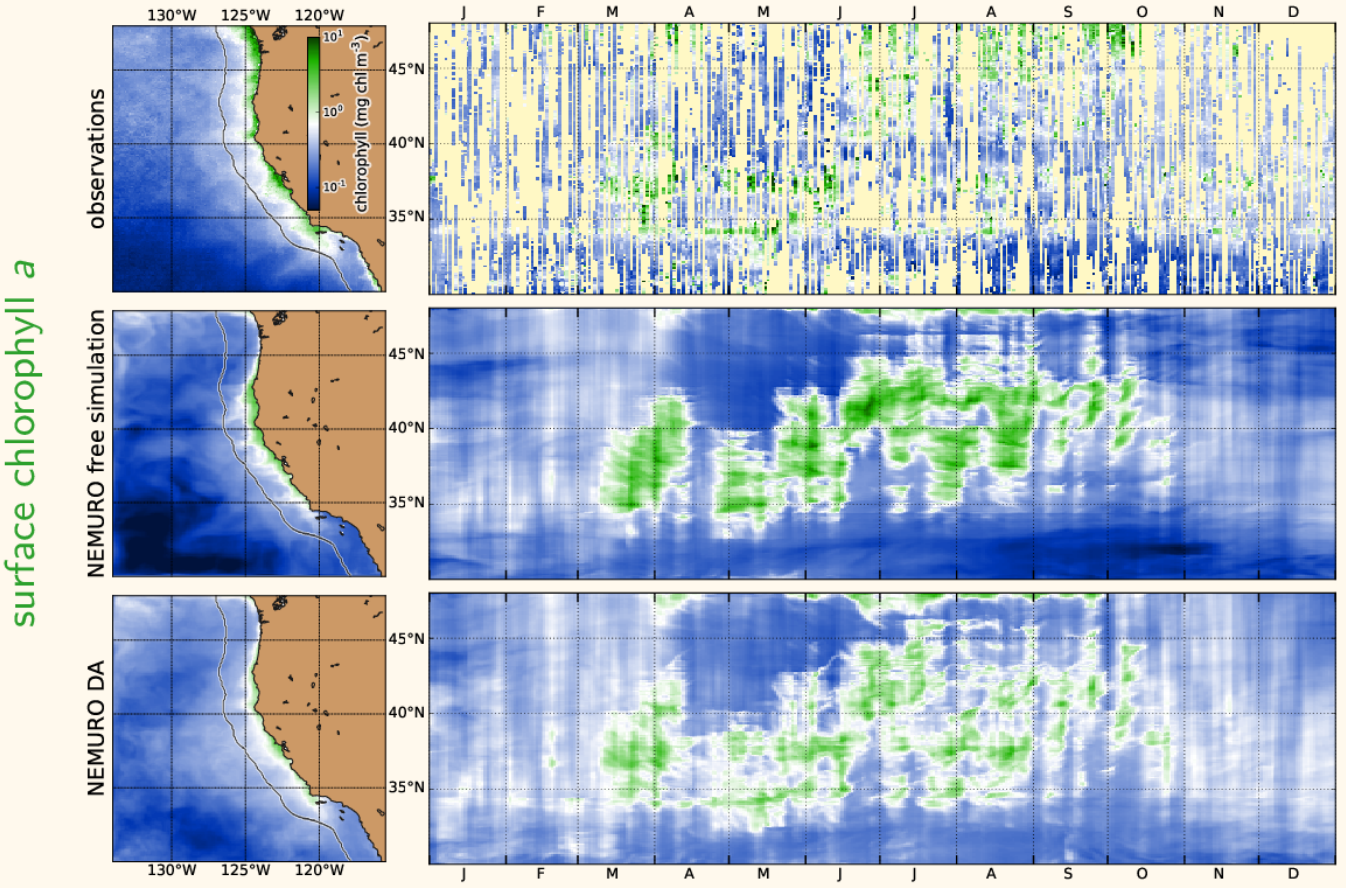
Ocean Sciences Department, University of California Santa Cruz

EuroSea workshop 2022



4dVar physical+biogeochemical data assimilation for the NEMURO model

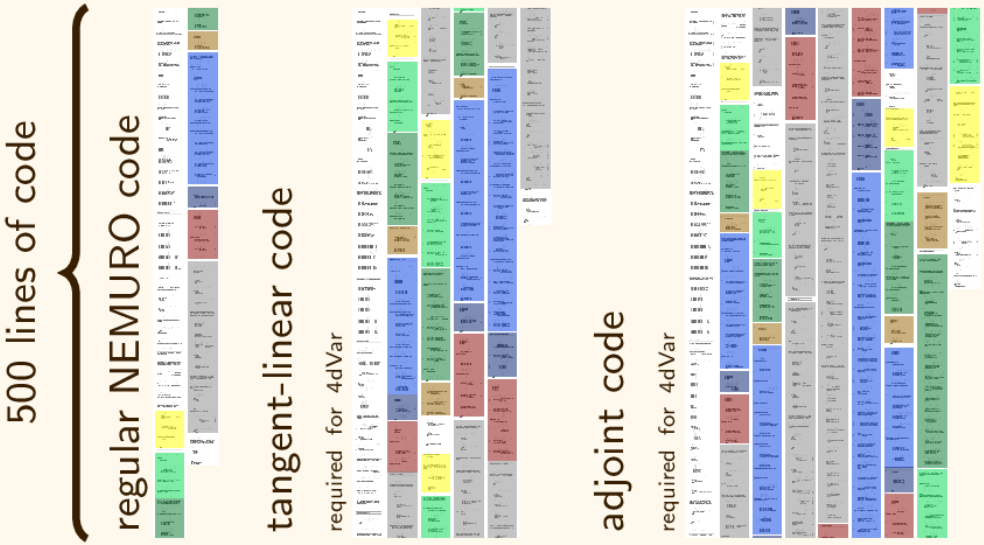
- We have successfully applied 4dVar-based state estimation to the medium complexity (11 variable) NEMURO model.



Mattern et al., OM (2017)

4dVar physical+biogeochemical data assimilation for the NEMURO model

- We have successfully applied 4dVar-based state estimation to the medium complexity (11 variable) NEMURO model.
- We would like to use state estimation for our regional implementation of the (37 variable) Darwin model.
- But 4dVar data assimilation requires tangent-linear and adjoint code which is difficult and cumbersome to create and maintain when the code is changed:



- The Darwin code is more complex and longer than the NEMURO code, and we do not want to create tangent-linear or adjoint code for Darwin by hand.

the 4dEnOI technique

We aimed to develop and implement a data assimilation technique that

1. does not require tangent-linear or adjoint code and is relatively easy to implement,
2. creates similarly good results as our benchmark 4dVar implementation, and
3. is fast enough computationally, to be used with the Darwin model.

The result is an ensemble optimal interpolation technique that is 4-dimensional ("4dEnOI") and uses an ensemble of non-assimilative simulations to compute flow-dependent statistics:

4dEnOI update:

$$\mathbf{x}^* = \mathbf{x} + (\alpha\,\mathbf{L} \circ \mathrm{cov}\,(\mathbf{X}, H(\mathbf{X}))) \left(\alpha\,\mathbf{L}' \circ \mathrm{cov}\,(H(\mathbf{X}), H(\mathbf{X}))) + \mathbf{R}\right)^{-1} (\mathbf{y} - H(\mathbf{x}))$$

the 4dEnOI technique

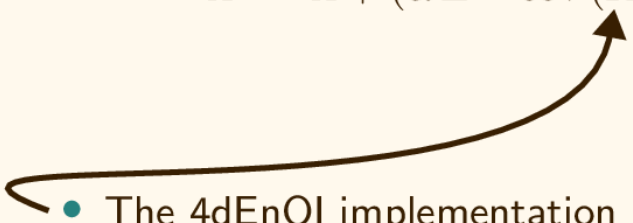
We aimed to develop and implement a data assimilation technique that

1. does not require tangent-linear or adjoint code and is relatively easy to implement,
2. creates similarly good results as our benchmark 4dVar implementation, and
3. is fast enough computationally, to be used with the Darwin model.

The result is an ensemble optimal interpolation technique that is 4-dimensional ("4dEnOI") and uses an ensemble of non-assimilative simulations to compute flow-dependent statistics:

4dEnOI update:

$$\mathbf{x}^* = \mathbf{x} + (\alpha\,\mathbf{L} \circ \mathrm{cov}\,(\mathbf{X}, H(\mathbf{X}))) \left(\alpha\,\mathbf{L}' \circ \mathrm{cov}\,(H(\mathbf{X}), H(\mathbf{X}))) + \mathbf{R}\right)^{-1} (\mathbf{y} - H(\mathbf{x}))$$



- The 4dEnOI implementation uses a background ensemble of non-assimilative simulations.

the 4dEnOI technique

We aimed to develop and implement a data assimilation technique that

1. does not require tangent-linear or adjoint code and is relatively easy to implement,
2. creates similarly good results as our benchmark 4dVar implementation, and
3. is fast enough computationally, to be used with the Darwin model.

The result is an ensemble optimal interpolation technique that is 4-dimensional ("4dEnOI") and uses an ensemble of non-assimilative simulations to compute flow-dependent statistics:

4dEnOI update:

$$\mathbf{x}^* = \mathbf{x} + (\alpha\,\mathbf{L} \circ \mathrm{cov}\,(\mathbf{X}, H(\mathbf{X}))) \left(\alpha\,\mathbf{L}' \circ \mathrm{cov}\,(H(\mathbf{X}), H(\mathbf{X}))) + \mathbf{R}\right)^{-1} (\mathbf{y} - H(\mathbf{x}))$$

- The 4dEnOI implementation uses a background ensemble of non-assimilative simulations.
- There is no linearization of the observation operator.

the 4dEnOI technique

We aimed to develop and implement a data assimilation technique that

1. does not require tangent-linear or adjoint code and is relatively easy to implement,
2. creates similarly good results as our benchmark 4dVar implementation, and
3. is fast enough computationally, to be used with the Darwin model.

The result is an ensemble optimal interpolation technique that is 4-dimensional ("4dEnOI") and uses an ensemble of non-assimilative simulations to compute flow-dependent statistics:

4dEnOI update:

$$\mathbf{x}^* = \mathbf{x} + (\alpha\,\mathbf{L} \circ \mathrm{cov}\,(\mathbf{X}, H(\mathbf{X}))) \left(\alpha\,\mathbf{L}' \circ \mathrm{cov}\,(H(\mathbf{X}), H(\mathbf{X}))) + \mathbf{R}\right)^{-1} (\mathbf{y} - H(\mathbf{x}))$$

- The 4dEnOI implementation uses a background ensemble of non-assimilative simulations.
- There is no linearization of the observation operator.
- cov is the sample (ensemble) covariance function.

the 4dEnOI technique

We aimed to develop and implement a data assimilation technique that

1. does not require tangent-linear or adjoint code and is relatively easy to implement,
2. creates similarly good results as our benchmark 4dVar implementation, and
3. is fast enough computationally, to be used with the Darwin model.

The result is an ensemble optimal interpolation technique that is 4-dimensional ("4dEnOI") and uses an ensemble of non-assimilative simulations to compute flow-dependent statistics:

4dEnOI update:

$$\mathbf{x}^* = \mathbf{x} + (\alpha\,\mathbf{L} \circ \mathrm{cov}\,(\mathbf{X}, H(\mathbf{X}))) \left(\alpha\,\mathbf{L}' \circ \mathrm{cov}\,(H(\mathbf{X}), H(\mathbf{X}))) + \mathbf{R}\right)^{-1} (\mathbf{y} - H(\mathbf{x}))$$

- The 4dEnOI implementation uses a background ensemble of non-assimilative simulations.
- There is no linearization of the observation operator.
- cov is the sample (ensemble) covariance function.
- $\mathbf{x}$ is the first ensemble member.

the 4dEnOI technique

We aimed to develop and implement a data assimilation technique that

1. does not require tangent-linear or adjoint code and is relatively easy to implement,
2. creates similarly good results as our benchmark 4dVar implementation, and
3. is fast enough computationally, to be used with the Darwin model.

The result is an ensemble optimal interpolation technique that is 4-dimensional ("4dEnOI") and uses an ensemble of non-assimilative simulations to compute flow-dependent statistics:

4dEnOI update:

$$\mathbf{x}^* = \mathbf{x} + (\alpha\,\mathbf{L} \circ \mathrm{cov}\,(\mathbf{X}, H(\mathbf{X}))) \left(\alpha\,\mathbf{L}' \circ \mathrm{cov}\,(H(\mathbf{X}), H(\mathbf{X}))) + \mathbf{R}\right)^{-1} (\mathbf{y} - H(\mathbf{x}))$$

- The 4dEnOI implementation uses a background ensemble of non-assimilative simulations.
- There is no linearization of the observation operator.
- cov is the sample (ensemble) covariance function.
- $\mathbf{x}$ is the first ensemble member.
- The matrices $\mathbf{L} \in \mathbb{R}^{n_{\text{state}} \times n_{\text{obs}}}$ and $\mathbf{L}' \in \mathbb{R}^{n_{\text{obs}} \times n_{\text{obs}}}$ are used for localization.

We aimed to develop and implement a data assimilation technique that

1. does not require tangent-linear or adjoint code and is relatively easy to implement,
2. creates similarly good results as our benchmark 4dVar implementation, and
3. is fast enough computationally, to be used with the Darwin model.

The result is an ensemble optimal interpolation technique that is 4-dimensional ("4dEnOI") and uses an ensemble of non-assimilative simulations to compute flow-dependent statistics:

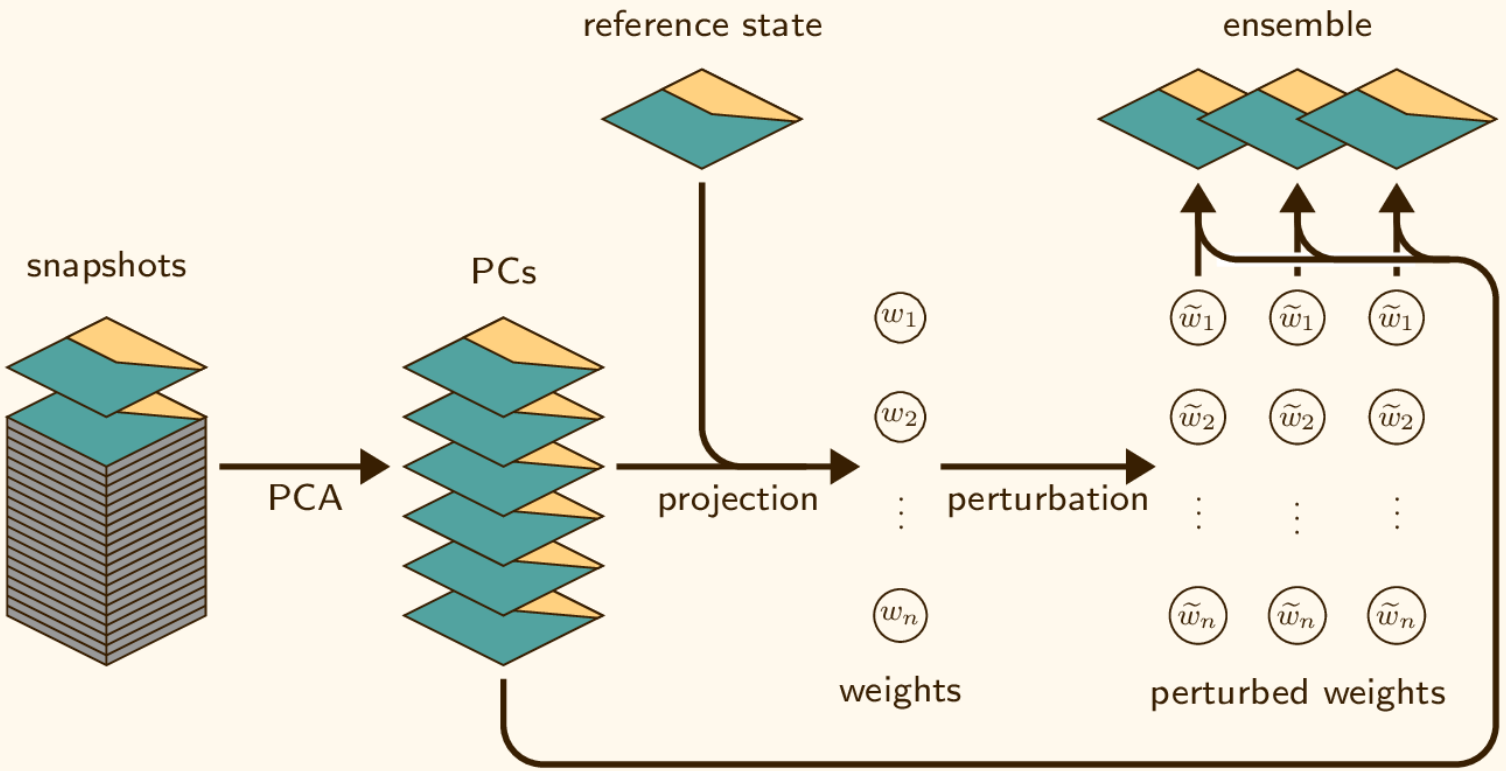
4dEnOI update:

$$\mathbf{x}^* = \mathbf{x} + \left(\alpha\,\mathbf{L} \circ \operatorname{cov}\left(\mathbf{X}, H(\mathbf{X})\right)\right)\left(\alpha\,\mathbf{L}' \circ \operatorname{cov}\left(H(\mathbf{X}), H(\mathbf{X})\right) + \mathbf{R}\right)^{-1}\left(\mathbf{y} - H(\mathbf{x})\right)$$

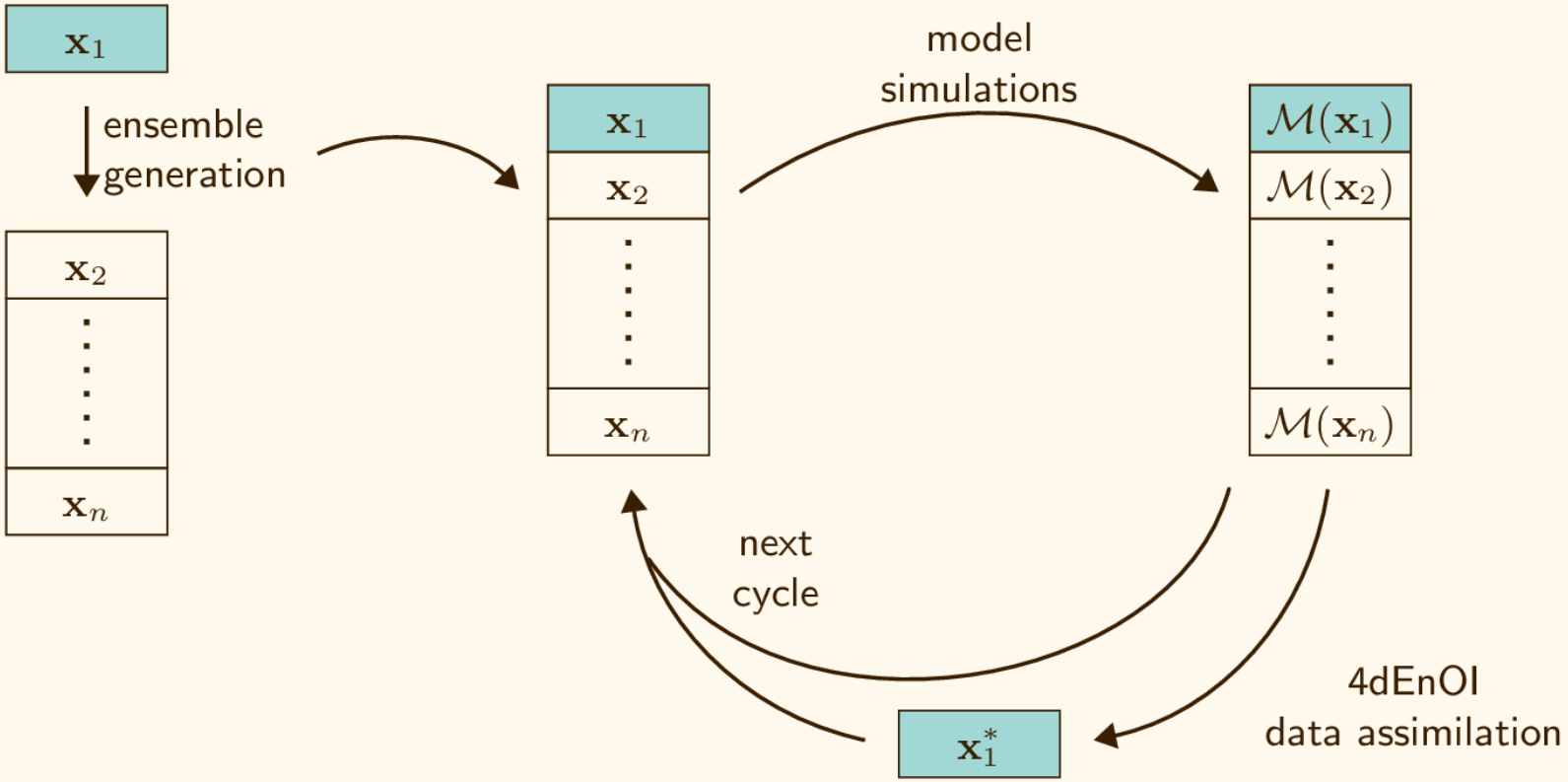
- The 4dEnOI implementation uses a background ensemble of non-assimilative simulations.
- There is no linearization of the observation operator.
- cov is the sample (ensemble) covariance function.
- $\mathbf{x}$ is the first ensemble member.
- The matrices $\mathbf{L} \in \mathbb{R}^{n_{\text{state}} \times n_{\text{obs}}}$ and $\mathbf{L}' \in \mathbb{R}^{n_{\text{obs}} \times n_{\text{obs}}}$ are used for localization.
- The scaling factor $\alpha \in]0, 1]$ may be used to reduce the increment. In previous EnOI studies, it was reduced from a value of 1 to account for the use of non time-evolving, static background error covariance matrices. In our reference implementation, $\alpha = 1$.

ensemble generation using a PCA

- At the start, the ensemble is generated from a reference state using snapshots from a long (non-assimilative) simulation and a principal component analysis (PCA).

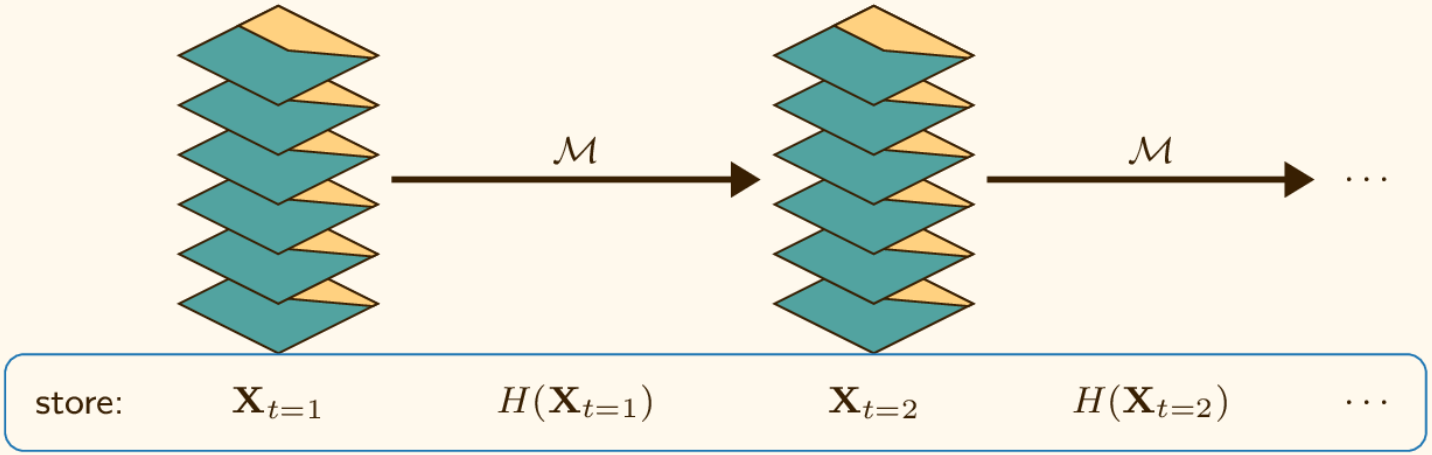


4dEnOI setup



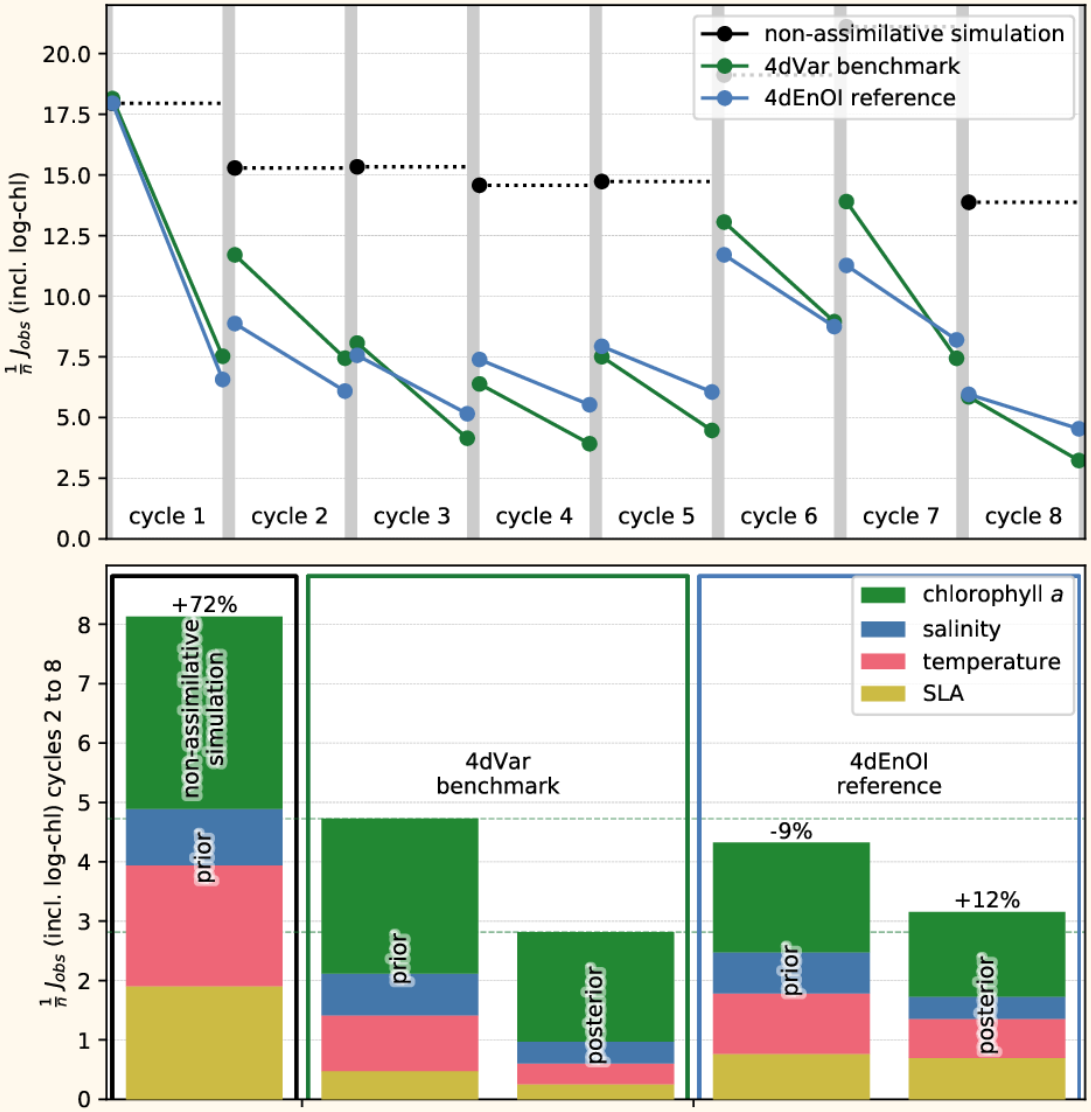
Why use 4dEnOI?

- We can use a static ensemble (not including the first ensemble member that is adjusted by the data assimilation):



- Updating a single ensemble member (including localization) is significantly cheaper than updating the full ensemble (as in the EnKF; factor of ≈ 1.8 in current implementation).

comparison to 4dVar benchmark



Test bed for the following data assimilation experiments:

- ROMS NPZD model (`NPZD_IRON`) in U.S. west coast domain
- 8 4-day cycles starting in April 2019 (time with active biology)
- real observations for sea level anomaly (SLA), temperature, salinity and chlorophyll *a* (converted to observations of phytoplankton)
- $\leq$30 000 observations in total per cycle (some thinning)
- 4dVar benchmark uses same initial conditions as the first ensemble member in 4dEnOI

The 4dEnOI reference implementation uses:

- an ensemble of 25 members
- localization: $\mathbf{L} = \mathbf{L}_{\mathsf{x}} \circ \mathbf{L}_{\mathsf{y}} \circ \mathbf{L}_{\mathsf{z}} \circ \mathbf{L}_{\mathsf{var}} \in [0,1]^{n_{\mathsf{state}} \times n_{\mathsf{obs}}}$
  - horizontal localization (length scale: 10 grid cells)
  - vertical localization (length scale: 300 m)
  - variable localization (localization strength: 0.3; not applied to physical-physical covariance entries)

variable localization implementation

- Variable localization reduces the influence of spurious correlations between different variables even in spatially close grid cells.
- Entries in the covariance matrices $\mathbf{L}$ and $\mathbf{L}'$ are multiplied by $\omega = 0.3$ if they are associated with two different variables, unless both variables are physical variables.
- Equivalent to defining a distance between variables:

$$d_{\mathsf{var}}(v, w) = \begin{cases} 0 & \text{if } v = w \text{ or } v, w \in \{\text{physical variables}\} \\ 1 & \text{otherwise} \end{cases}$$

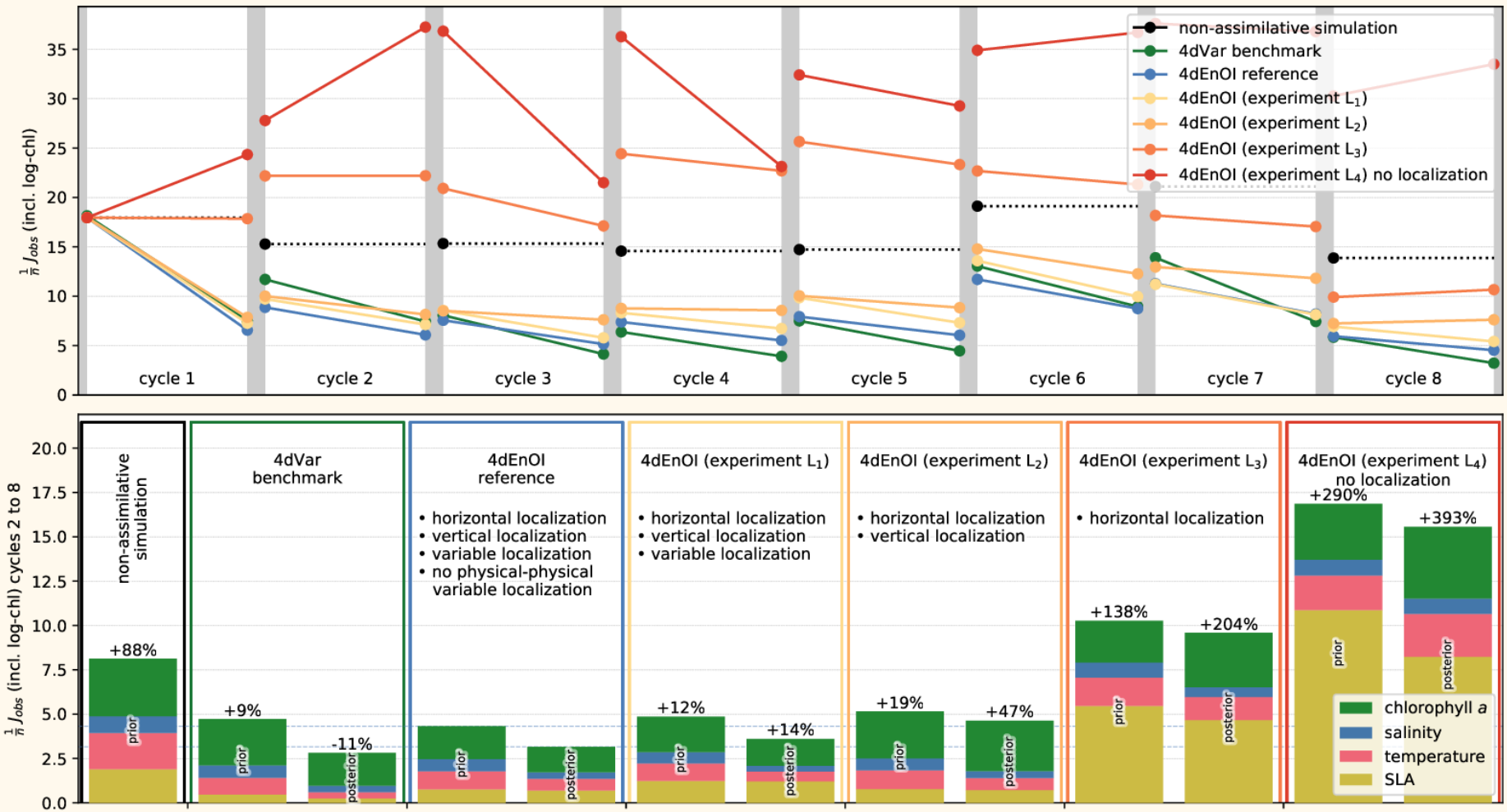
$$(\mathbf{L}_{\mathsf{var}})_{ij} = \omega^{d_{\mathsf{var}}(v_i, v_j)}$$

With adjustable weight ω set to 0.3.

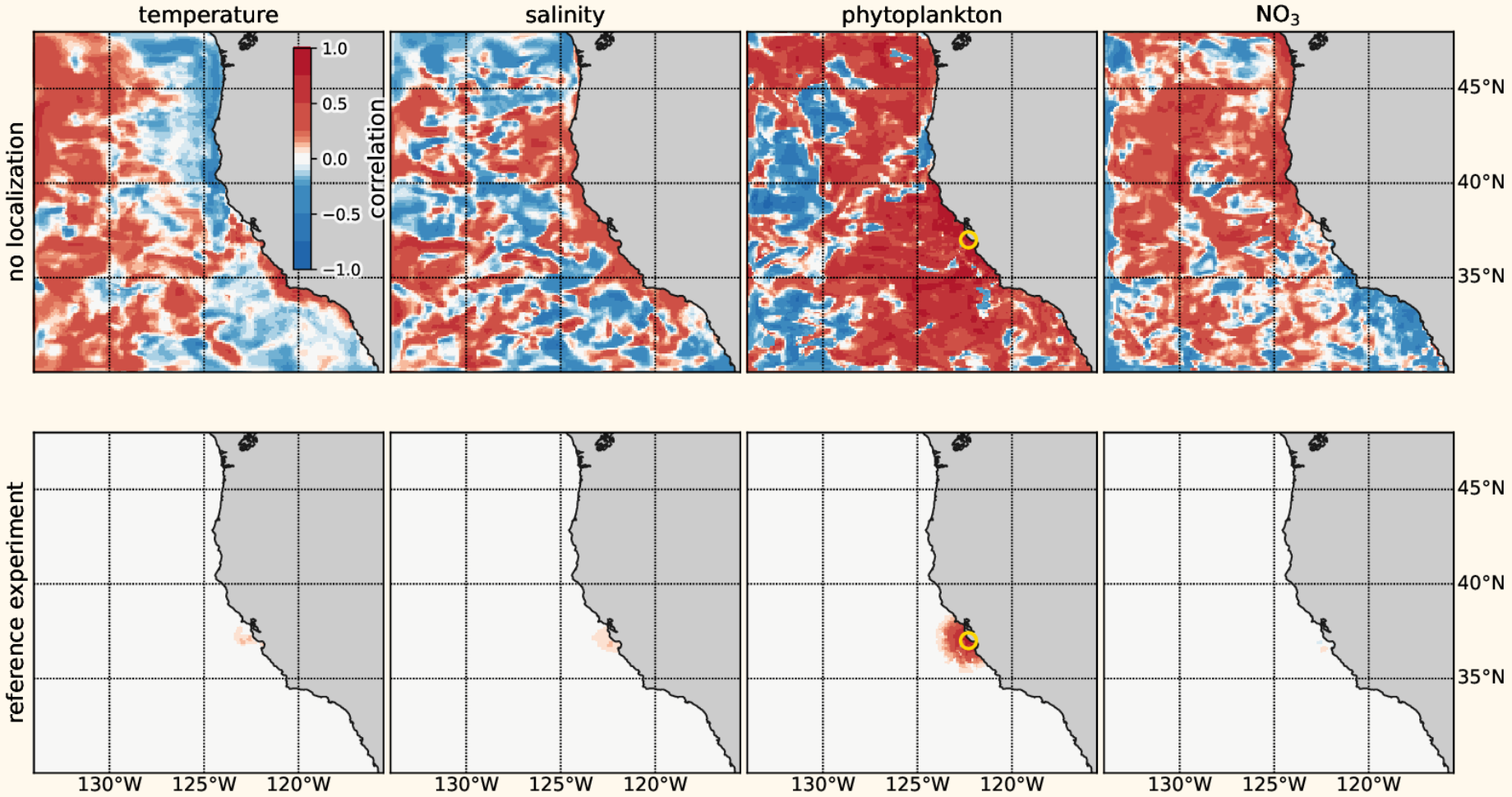
- We did not explore more complex approaches, for example consider biological model dynamics when defining d_{var}, for example:

$$0 < d_{\mathsf{var}}(\text{nutrients}, \text{phytoplankton}) < d_{\mathsf{var}}(\text{nutrients}, \text{zooplankton})$$

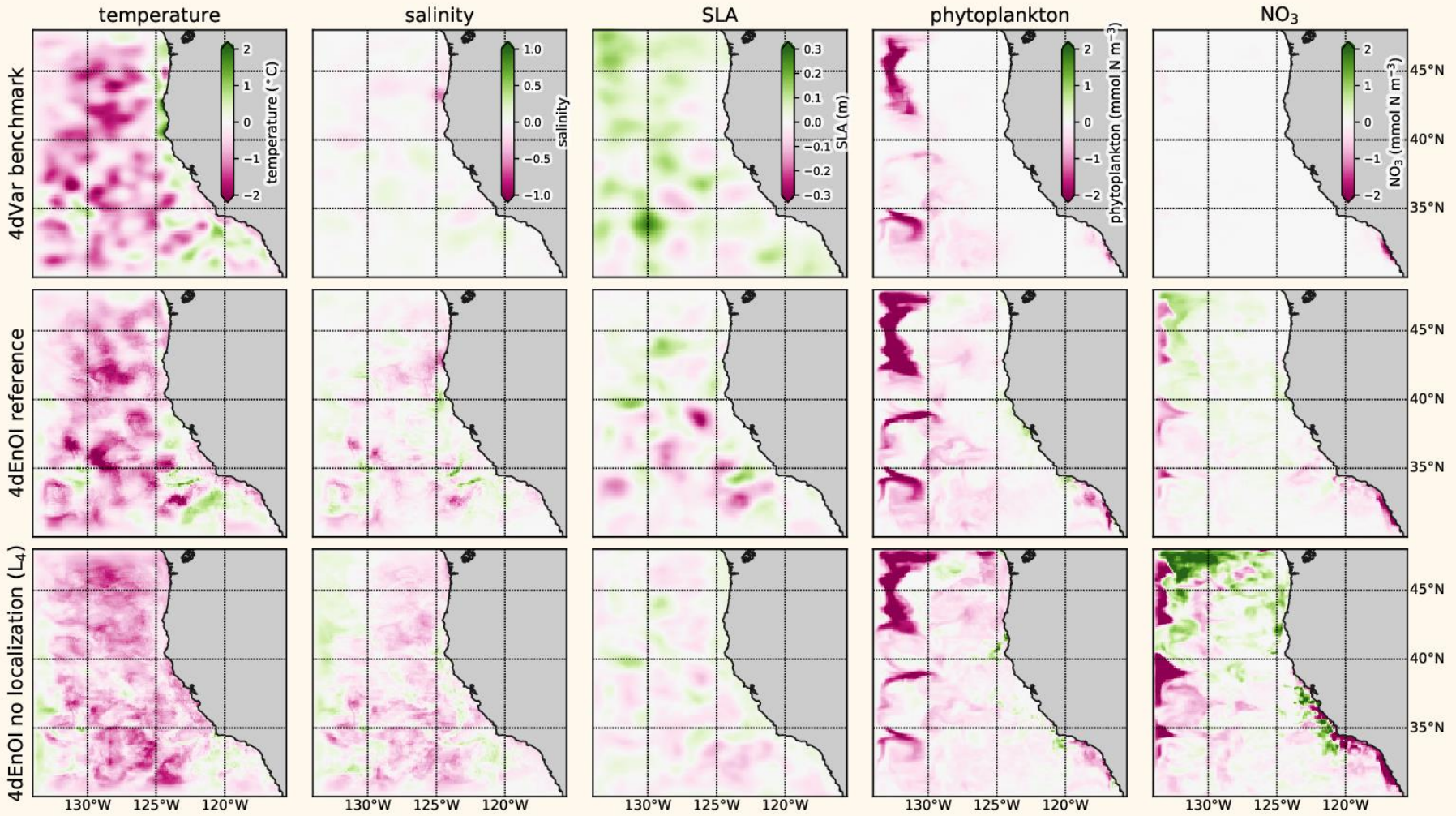
the effect of localization



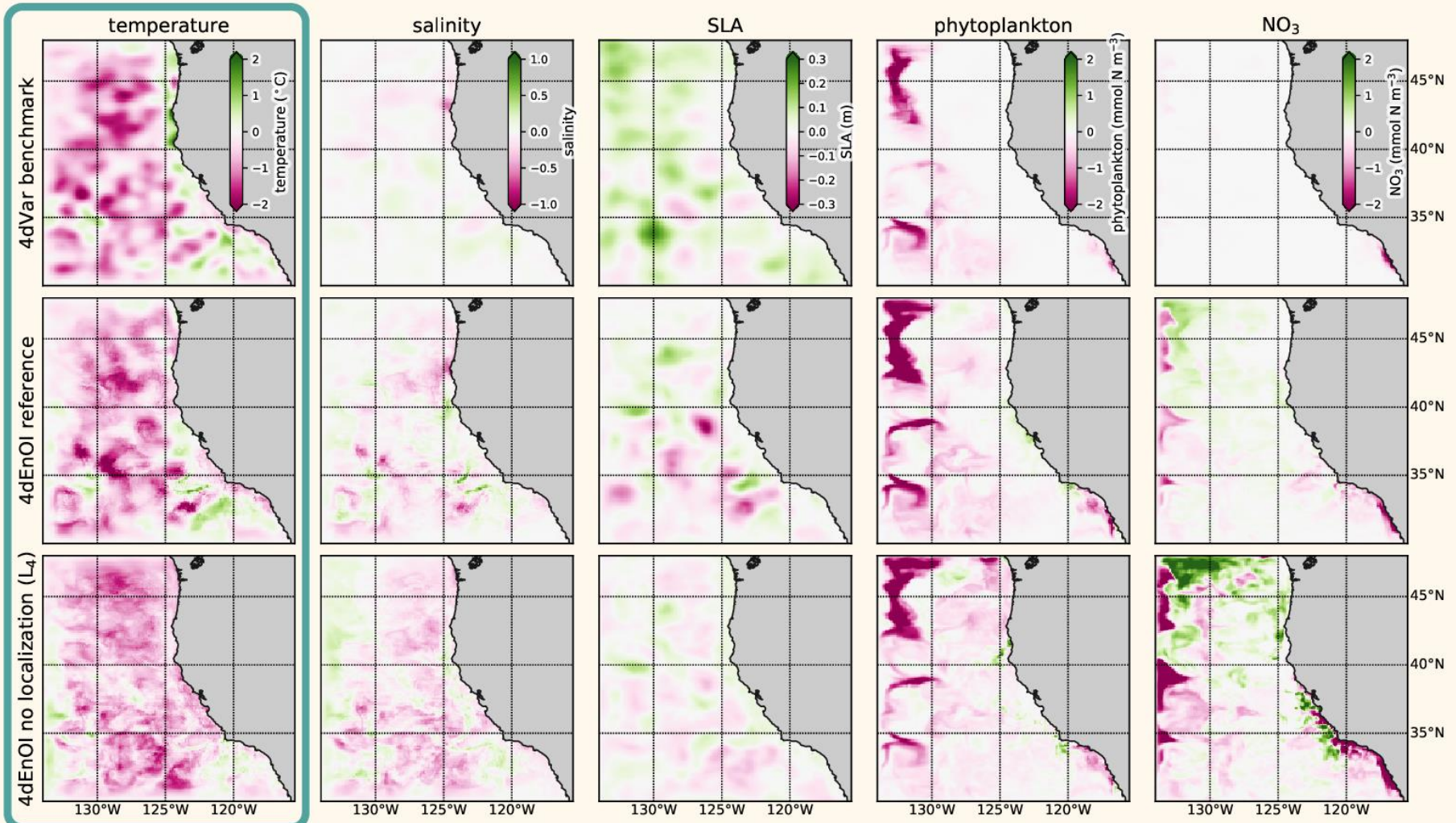
comparison of surface correlations (25 ensemble members)



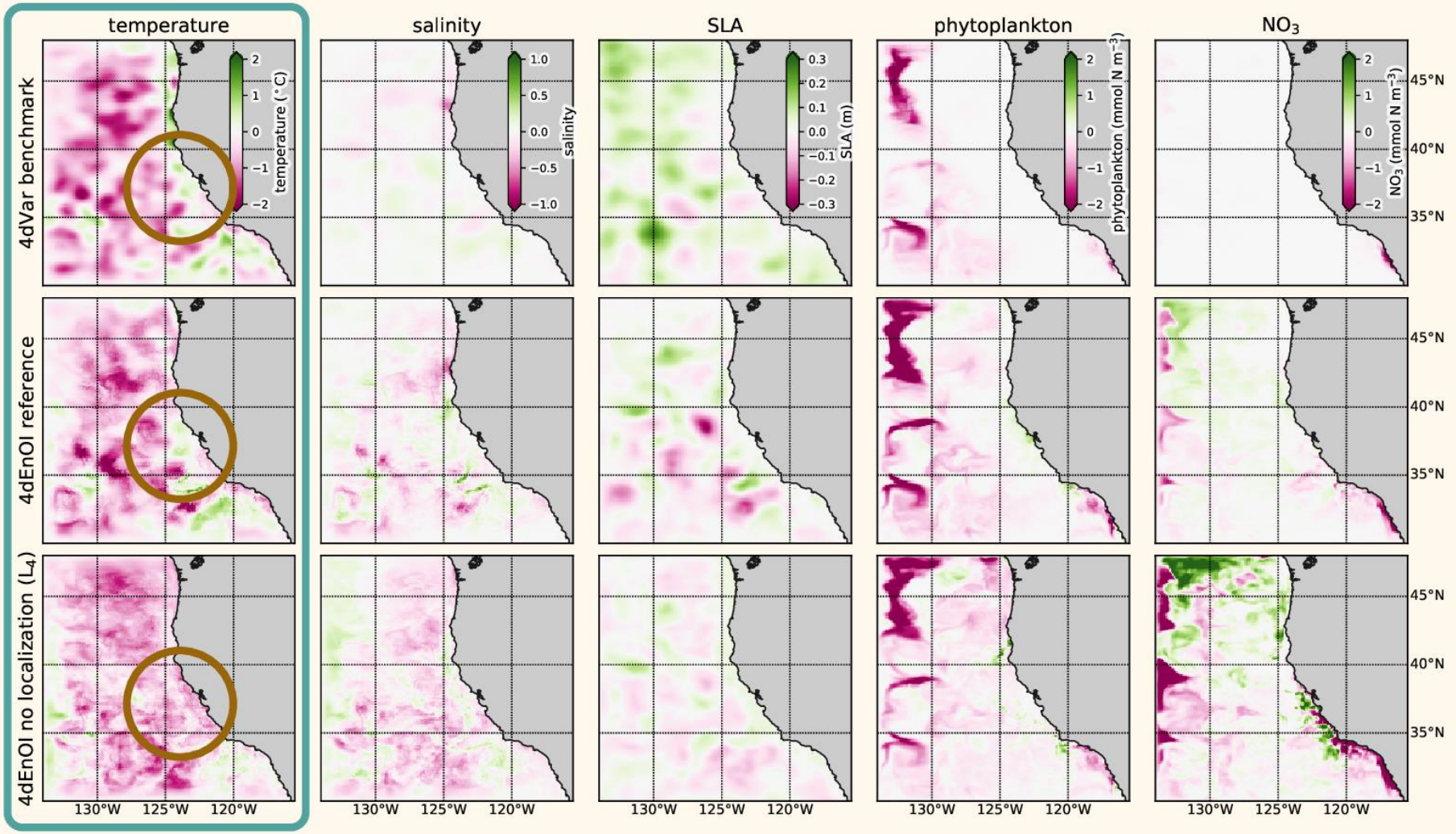
comparison to 4dVar benchmark – surface increments



comparison to 4dVar benchmark – surface increments



comparison to 4dVar benchmark – surface increments



summary

We developed a data assimilation technique that ...

1. ... does not require tangent linear or adjoint code and is relatively easy to implement. ✓
2. ... creates similarly good results as our benchmark 4dVar implementation. ✓
3. ... is fast enough computationally, to be used with the Darwin model. ✓
   - Computer runtime increases linearly with number of variables (still lots of room for optimization and especially parallelizing code).
   - Memory usage remains constant with an increasing number of variables.

future steps:

- The 4dEnOI framework is ready to be applied to the Darwin model.
- There are new challenges associated with Darwin state estimation, such as how to update $\geq$6 phytoplankton variables with chlorophyll *a* data.

We developed a data assimilation technique that ...

1. ... does not require tangent linear or adjoint code and is relatively easy to implement. ✓
2. ... creates similarly good results as our benchmark 4dVar implementation. ✓
3. ... is fast enough computationally, to be used with the Darwin model. ✓
   - Computer runtime increases linearly with number of variables (still lots of room for optimization and especially parallelizing code).
   - Memory usage remains constant with an increasing number of variables.

Thanks!

future steps:

- The 4dEnOI framework is ready to be applied to the Darwin model.
- There are new challenges associated with Darwin state estimation, such as how to update $\geq$6 phytoplankton variables with chlorophyll *a* data.