# Ensemble analysis and forecast of ecosystem indicators in the North Atlantic using ocean colour observations and prior statistics from a stochastic NEMO/PISCES simulator

**Jean-Michel Brankart, Mikhail Popov, Arthur Capet**
**Emmanuel Cosme, Pierre Brasseur**

IGE/CNRS
Grenoble

**SEAMLESS EU-H2020 project**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

## Context: H2020 **SEAMLESS** project (2021-2023) (www.seamlessproject.org)

- **Tier-3 project,** to sustain the evolution of the **« green »** component of the CMEMS, targetting impact at a 3-5 year horizon
- **WP3:** « Ensemble generation and assimilation methods »
- **Partners:** PML, NERSC, AWI, OGS, IGE

## The model configuration is based on **NEMO/PISCES**, as provided by **Mercator Ocean**

→ global coupled configuration, at 1/4° resolution

→ with explicit simulation of **model uncertainties**
(in link with the **Odessa** tier-2 project,
lead by the University of Liège)

# Towards a **simplified** analysis and forecasting system based on a prior ensemble model simulation

**Approach:**

→ Perform a **prior ensemble simulation**, with a state-of-the-art coupled circulation/ecosystem model

→ **Condition** this 4D ensemble **on ocean colour observations** to obtain the ensemble analysis and forecast

**Features:**

→ **Decouple** the complex models **simulations and** the **inversion** problem → more flexibility in the system

→ Focus the 4D inversion on **specific variables** on a **specific region and time window**

→ Directly estimate **ecosystem indicators**

**Shortcomings:**

→ The complex model is not used as a direct constraint in the inverse problem, but only indirectly through the prior ensemble

→ Prior uncertainty may be larger as compared to sequential systems

# Inverse problem

We **focus** on a small **4D subregion**
in the **North-East Atlantic** (31°-21°W, 44°-50.5°N):

40 X 40 grid points (10° x 7°, at 1/4° resolution)
X 32 levels (down to 220 m)
X 93 days (March 15 to June 15, 2019)
X 6 tracers (+ two 2D indicators) = **~ 30 x $10^6$** variables

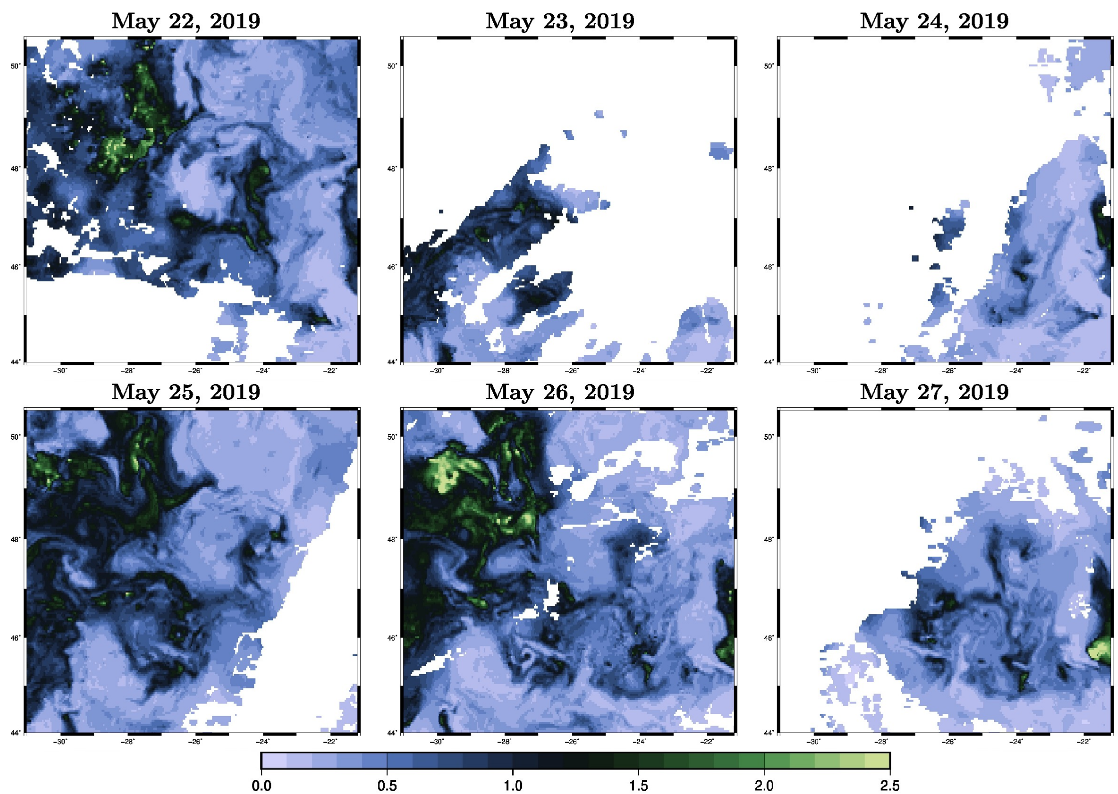
**Observation system:**

L3 chlorophyll product

between March 15
and June 15, 2019

Obs. error std: 30%

**~ 2 x $10^5$** observations



# The prior ensemble simulation

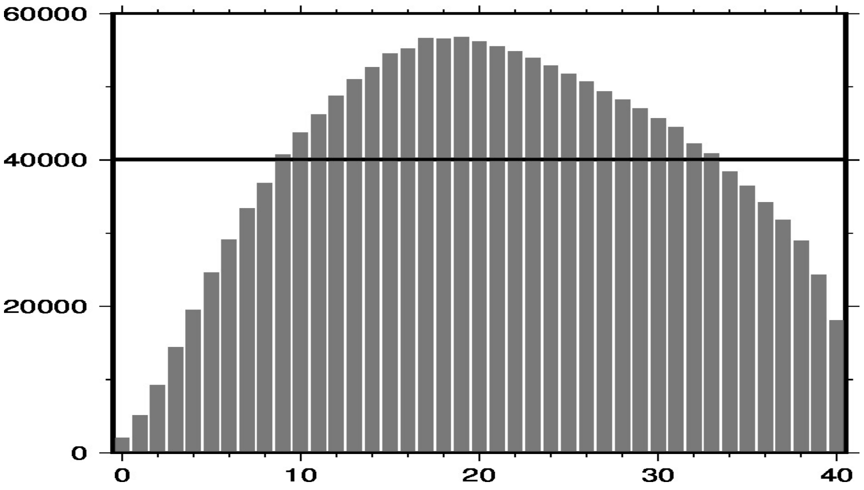
**A sample from the global NEMO/PISCES stochastic simulator:**

→ 40 ensemble members
→ daily outputs for our region of interest
→ embedding location uncertainties, parameter uncertainties, and uncertainties emerging from the unresolved scales

**Probabilistic scores have been applied to evaluate this ensemble simulation using L3 ocean colour observations.**

→ Example of rank histogram for the subregion used as an example below

→ In the North Atlantic Drift: 31°W-21°W, 44°N-50.5°N, April 21 to June 19, 2019

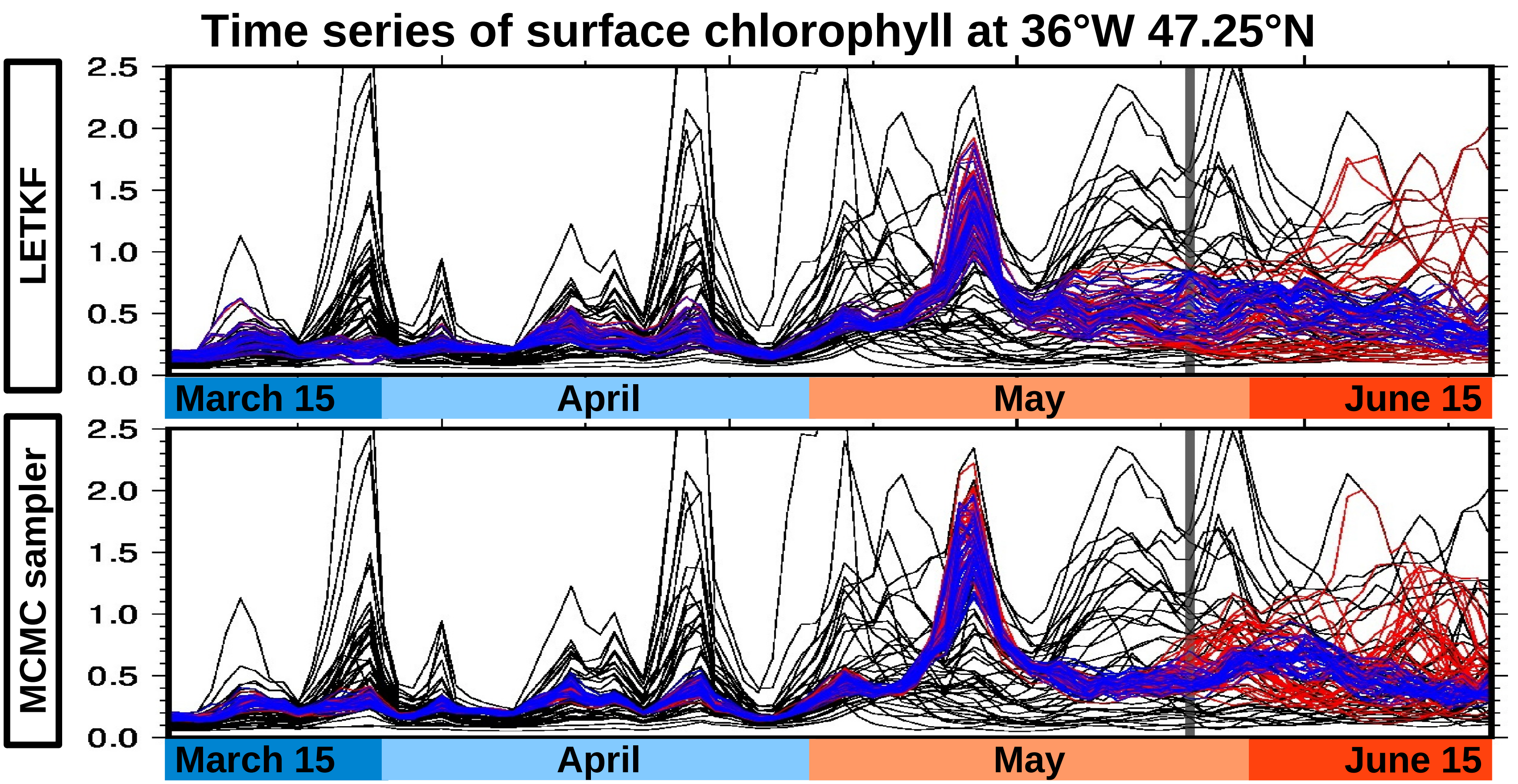



**The reliability of the prior ensemble is essential to obtain a reliable solution to the inverse problem**

# Methods : a brief summary

| | LETKF analysis | MCMC sampler |
|---|---|---|
| **Assumptions** | Gaussian background errors<br>Linear and local observation operator **H**<br>Gaussian observation error with prescribed covariance **R**<br>→ **Gaussian analysis errors** | Gaussian background errors<br>Nonlinear/nonlocal observation operator **H**<br>Non-Gaussian observation error conditioned on x: $p(y^o\|x)$.<br>→ **non-Gaussian analysis errors** |
| **Features** | domain localization<br>anamorphosis must be applied to x, $y^o$ and **R** | covariance localization<br>anamorphosis must be applied to x only |
| **In practice** | Linear algebra formula to compute Kalman gain<br>Loop and parallelization on local subdomains | Iterative method (MH algorithm), with fast random draws from the proposal distribution<br>Direct parallelization, with distributed x and $y^o$ |

# Results from the LETKF and MCMC sampler: ensemble analysis and forecast




**In black: prior ensemble simulations from NEMO/PISCES**

**In blue: ensemble analysis using all L3 observations**

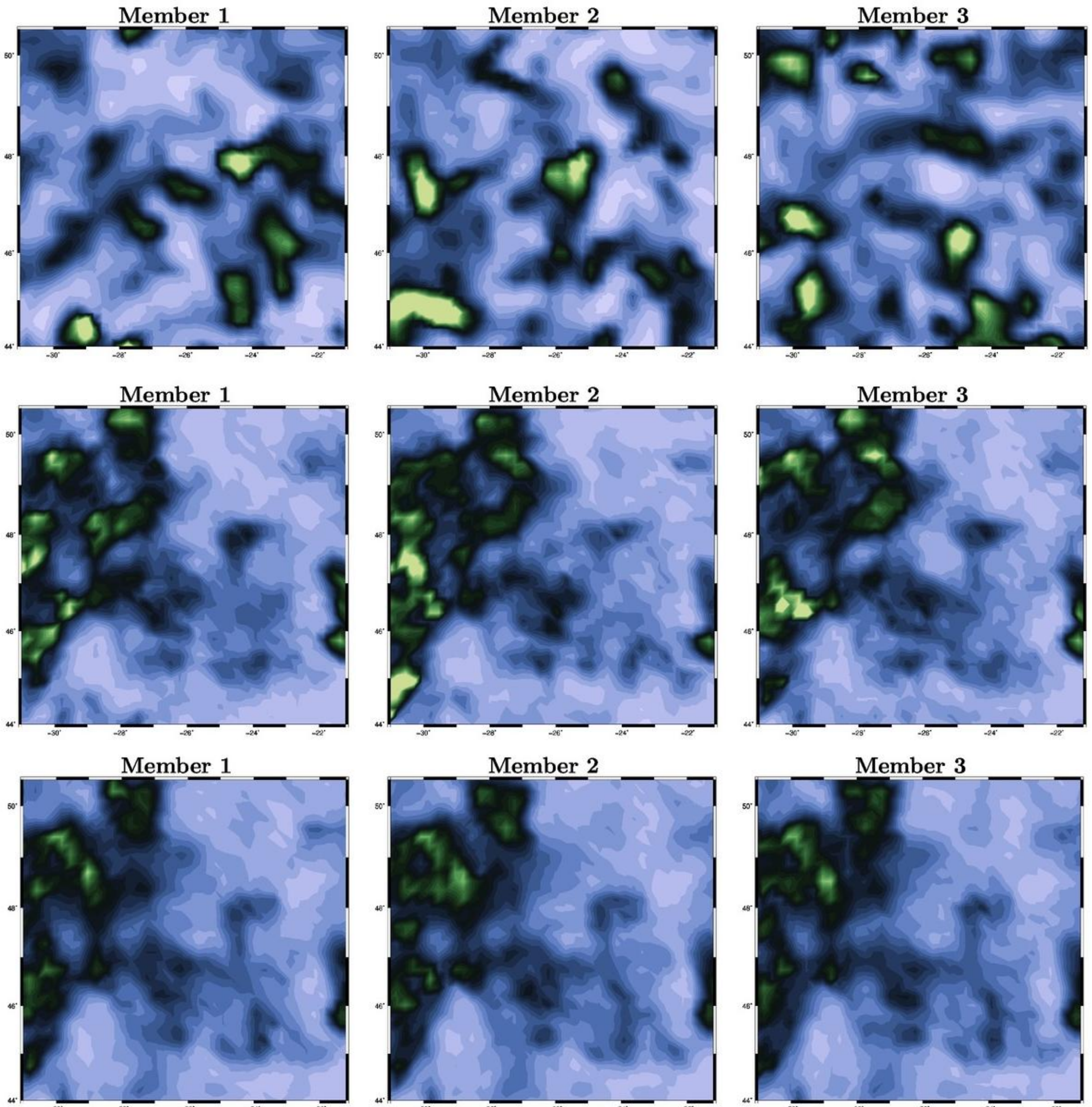
**In red: ensemble analysis and forecast using L3 observations until May 25 → some forecast skill for a few days**

# Results from the LETKF and MCMC sampler: **members of ensemble analysis for May 26, 2019** (using past and future observations)

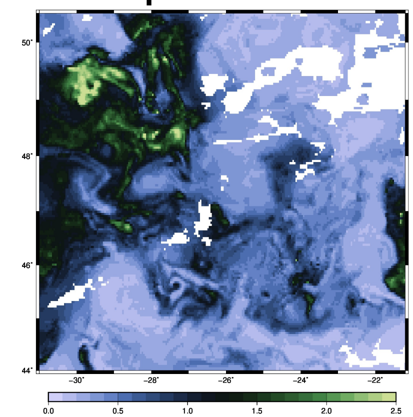
PRIOR

LETKF

MCMC sampler



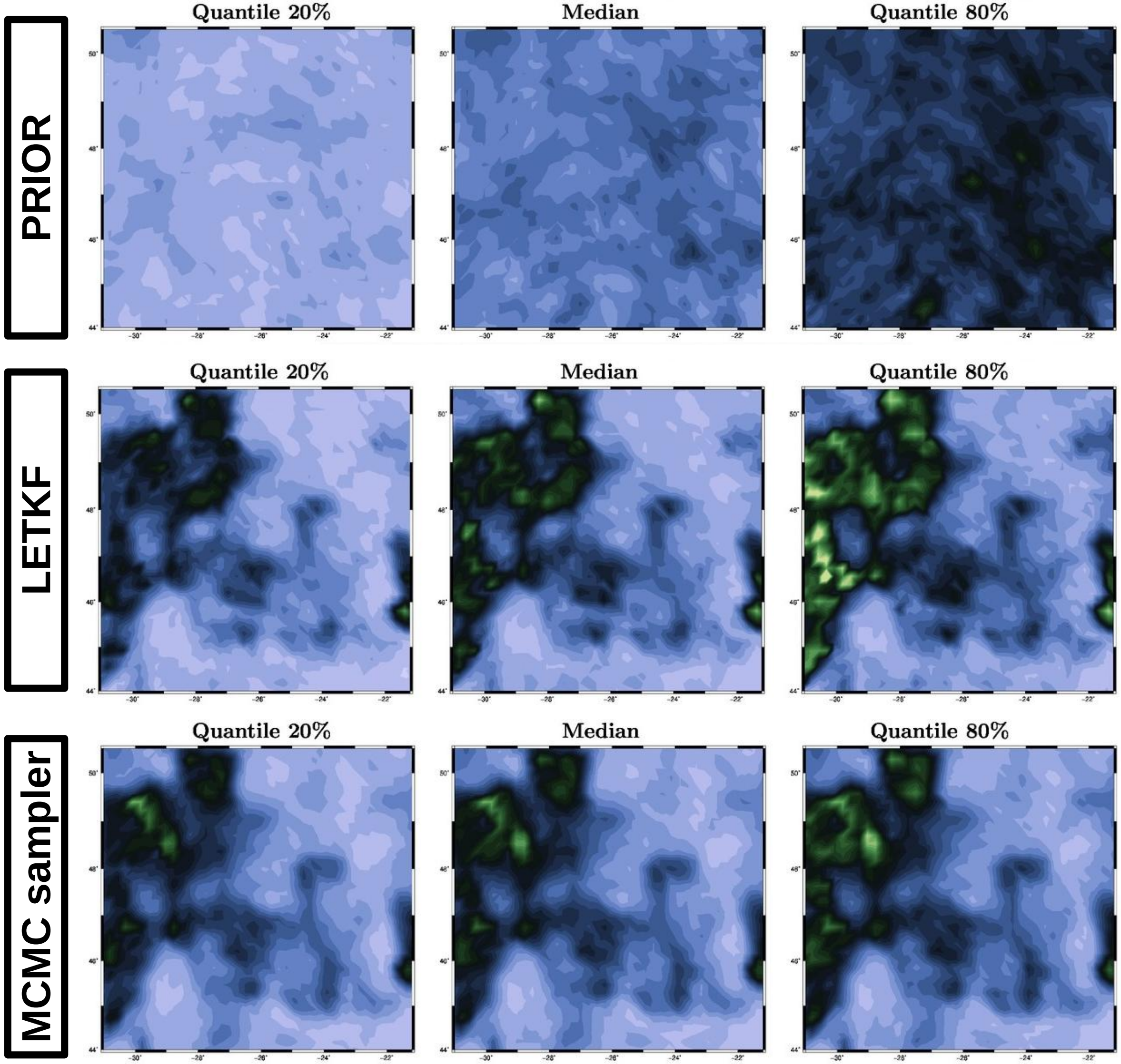

L3 chlorophyll product



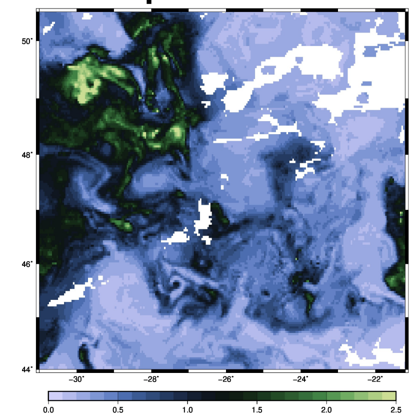

CRPS score
(with cross-validation)

Prior: 0.252 mg/m3
LETKF: 0.103 mg/m3
MCMC: 0.101 mg/m3

# Results from the LETKF and MCMC sampler: **<u>quantiles</u> of ensemble analysis for May 26, 2019** (using past and future observations)



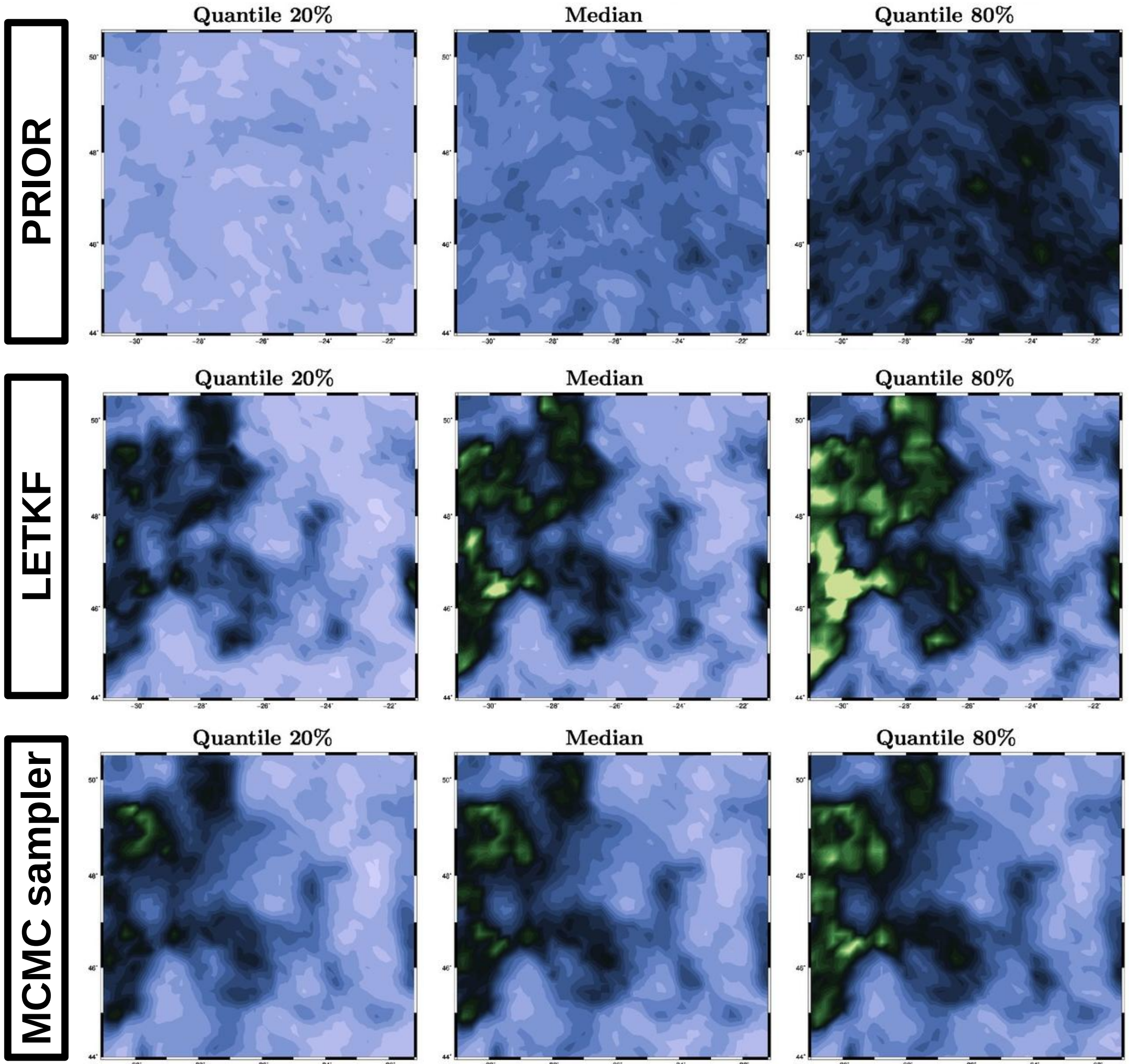

L3 chlorophyll product



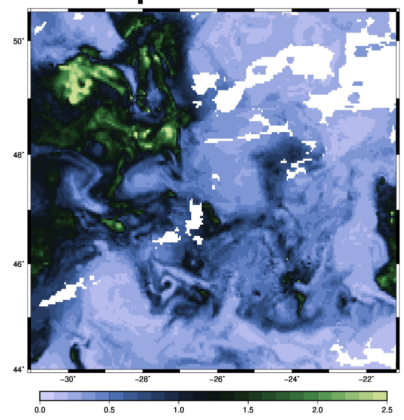

CRPS score
(with cross-validation)

Prior: 0.252 mg/m3
LETKF: 0.103 mg/m3
MCMC: 0.101 mg/m3

# Results from the LETKF and MCMC sampler:
# **1-day ensemble forecast for May 26, 2019**
# (i.e. using observations until May 25 only)



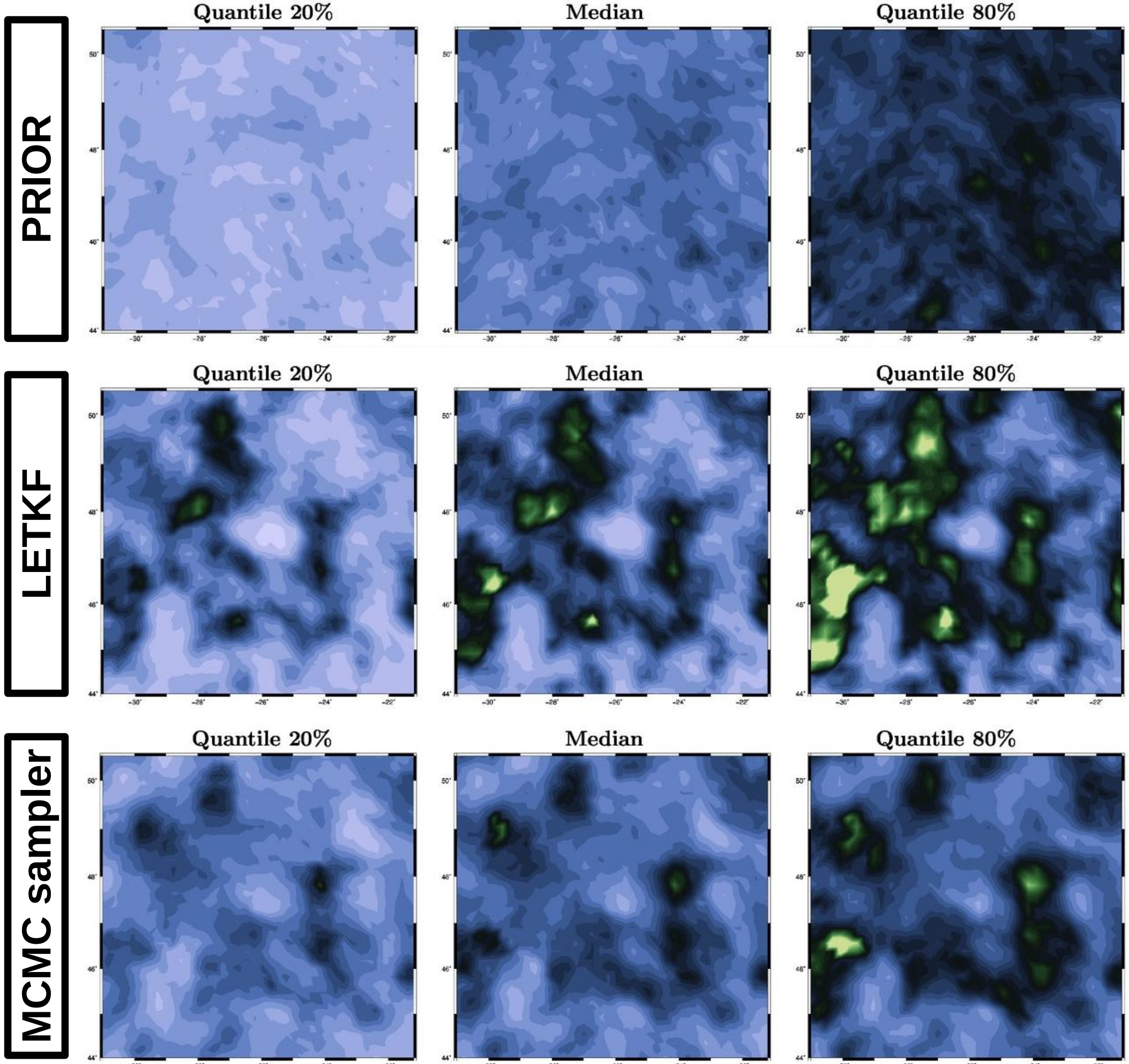

L3 chlorophyll product



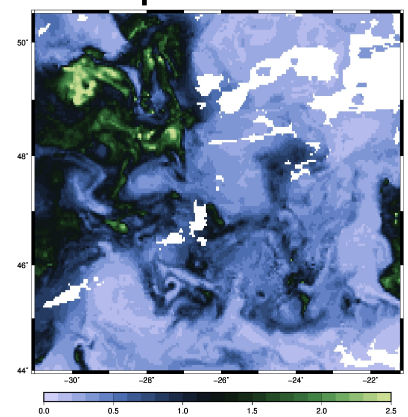

CRPS score

Prior: 0.252 mg/m3
LETKF: 0.134 mg/m3
MCMC: 0.132 mg/m3

# Results from the LETKF and MCMC sampler:
# **4-day ensemble forecast for May 26, 2019**
# (i.e. using observations until May 22 only)




L3 chlorophyll product

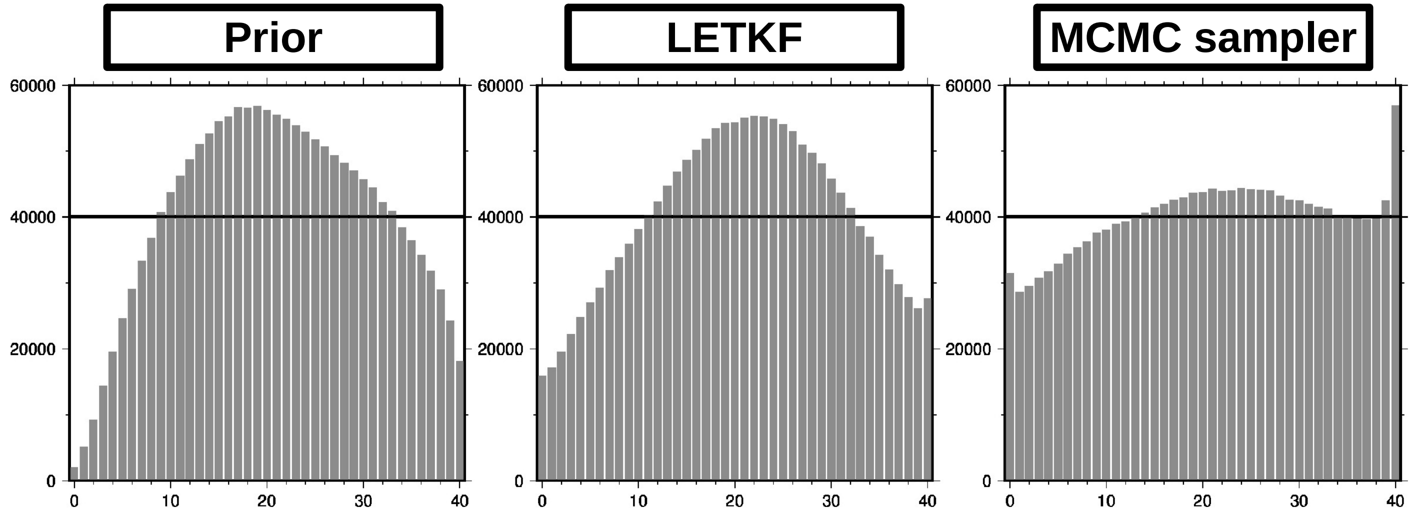



CRPS score

Prior: 0.252 mg/m3
LETKF: 0.194 mg/m3
MCMC: 0.214 mg/m3

# Summary of the probabilistic scores

<u>Rank Histograms</u> for the full time period

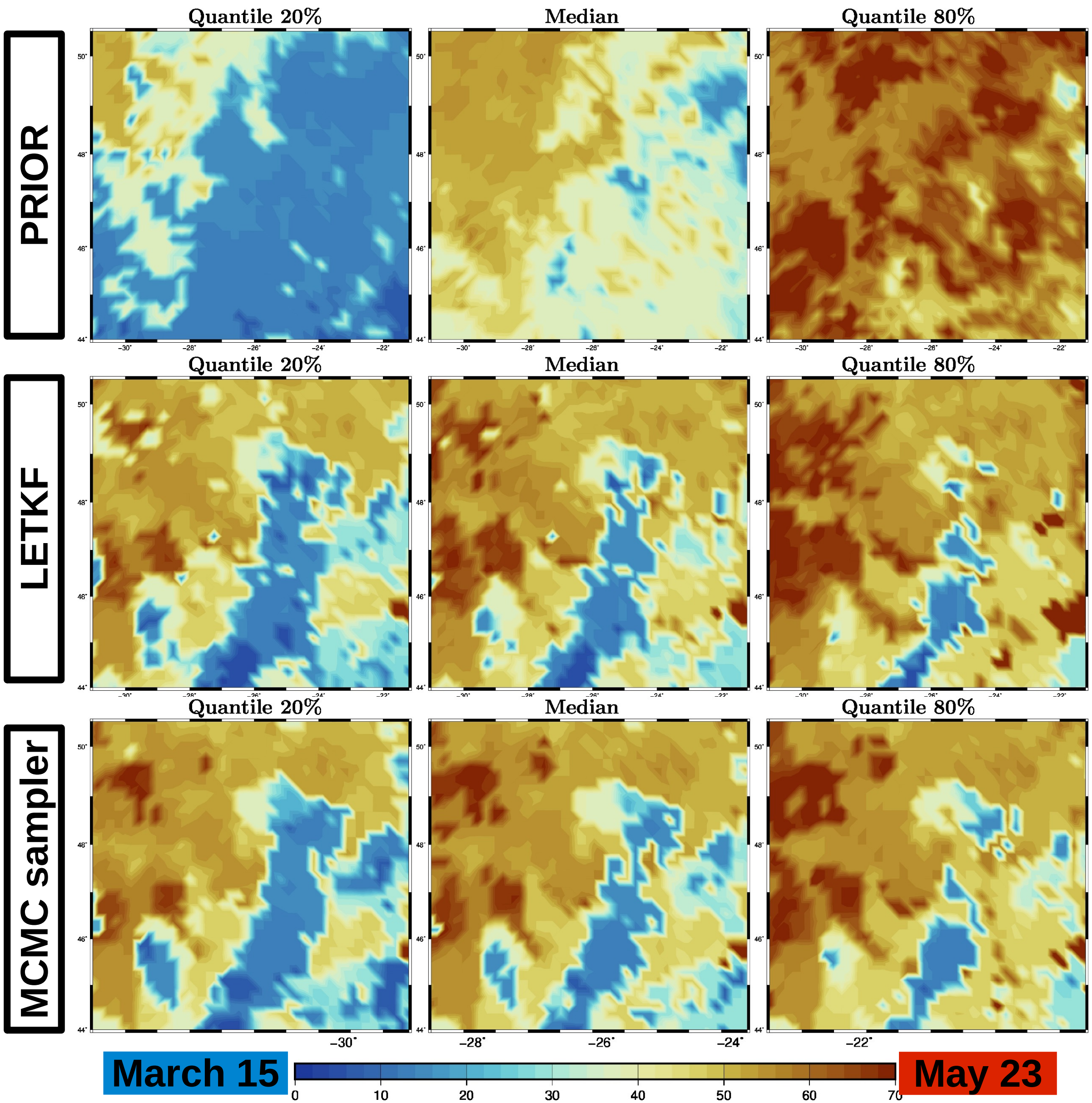



<u>CRPS and RCRV scores</u> for the analysis and forecast (May 26, 2019)

| Method | Experiment | CRPS Reliability | CRPS Resolution | RCRV Bias | RCRV Spread |
|---|---|---|---|---|---|
| Prior ensemble | | $1.0 \times 10^{-3}$ | 0.251 | 4.8 | 107.6 |
| LETKF algorithm | analysis | $5.8 \times 10^{-3}$ | 0.098 | 6.3 | 78.8 |
| | 1-day forecast | $4.9 \times 10^{-3}$ | 0.129 | -0.8 | 100.3 |
| | 4-day forecast | $9.8 \times 10^{-3}$ | 0.184 | -16.1 | 113.2 |
| | biased analysis | $13.4 \times 10^{-3}$ | 0.100 | 34.3 | 88.2 |
| MCMC sampler | analysis | $1.6 \times 10^{-3}$ | 0.099 | 7.6 | 102.6 |
| | 1-day forecast | $1.0 \times 10^{-3}$ | 0.131 | 3.8 | 134.1 |
| | 4-day forecast | $9.5 \times 10^{-3}$ | 0.204 | -8.3 | 165.1 |

Cross-validation is used to keep the observations independent in the computation of the analysis scores

# Ecosystem indicator: phenology for chlorophyll concentration



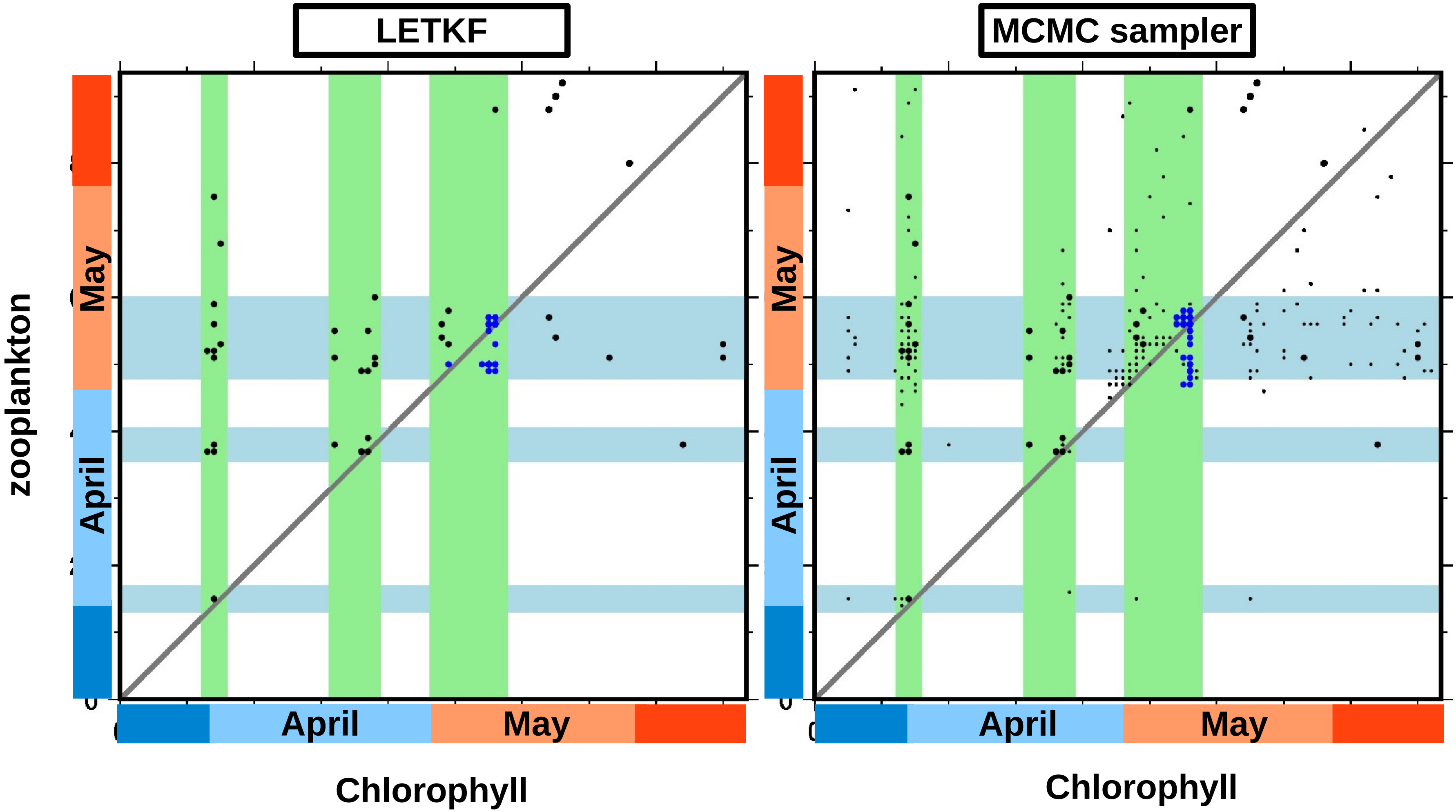

<u>A simple definition:</u>
first day at which the chlorophyll concentration reaches half of its maximum

The prior uncertainty is quite large

Strongly reduced by the observations

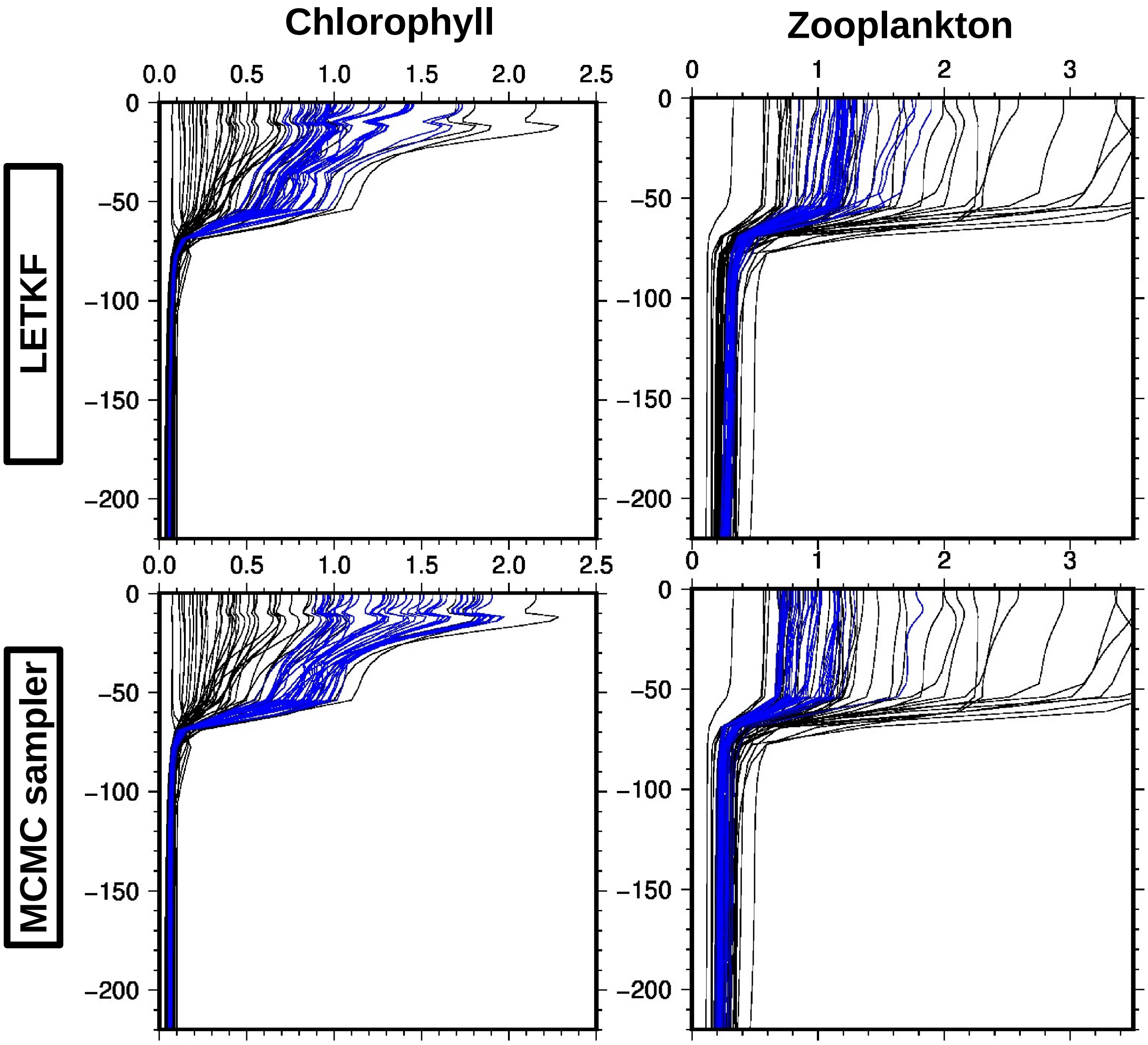
Quite consistently in the two methods

# Phenology for chlorophyll and zooplankton

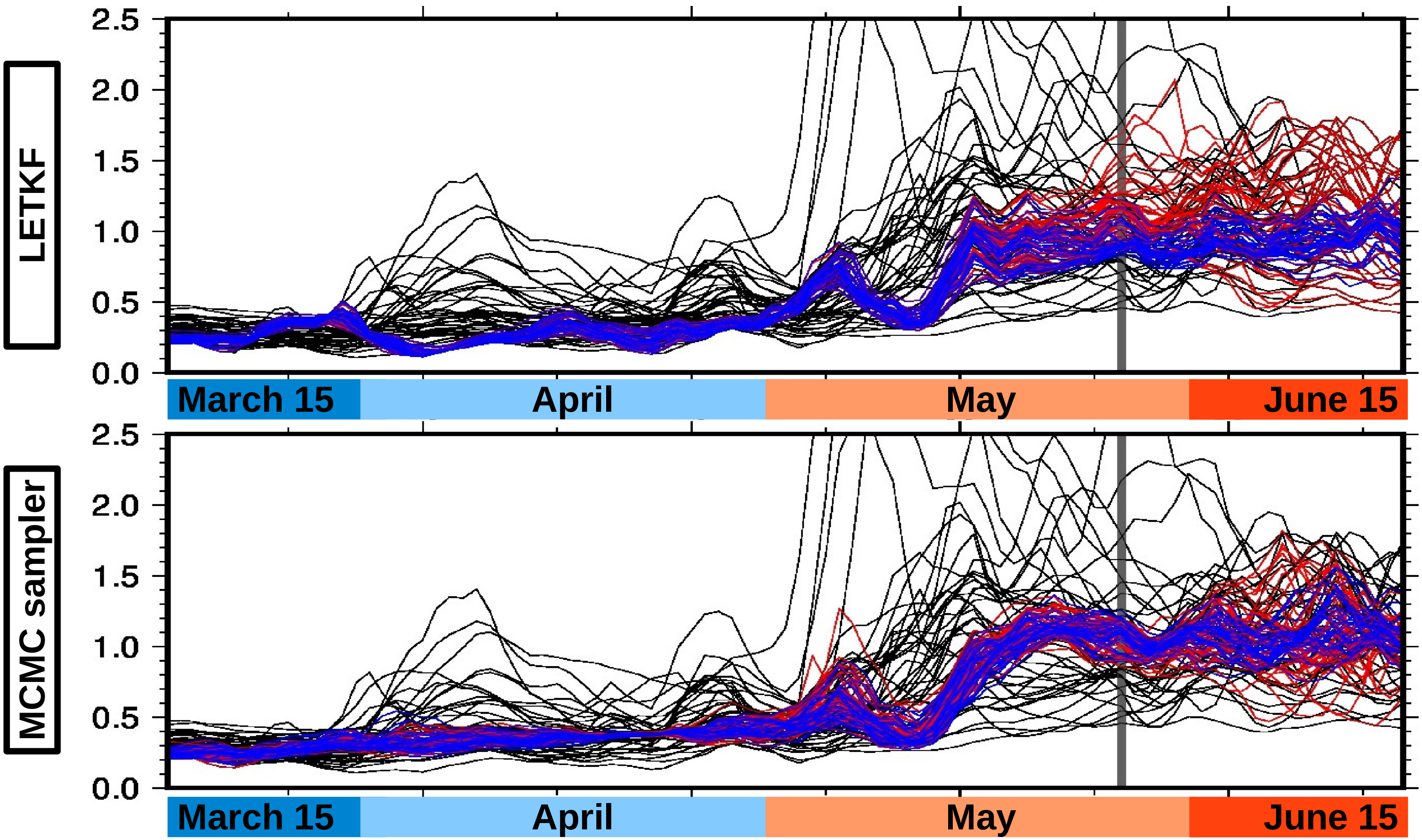



Three possible time windows in the prior ensemble.
Reduced to one in the posterior ensembles.

# Vertical profiles of chlorophyll and zooplankton



# Ecosystem indicator: trophic efficiency




**Time series of vertically integrated trophic efficiency at the center of the region of interest 36°W 47.25°N**

# Conclusions

In this example, the two methods produce **similar results**,
thus somehow validating each other.
In both cases, they produce an **ensemble analysis** that is continuous
in time, with a reliable description of posterior uncertainty,
at least for the observed variable.
The statistical **ensemble forecast** can contain valuable information
for a few days after the last observation.
For the **ecosystem indicators** (and any non-observed variable),
the reliability of the ensemble analysis and forecast
depends on the **reliability of the prior ensemble**.

A practical method to perform **4D ensemble analyses and forecasts.**

→ No need for full controllability of the complex model
(with so many state variables, when so few are observed).
No model restart from the analysis.

The focus is on **sampling possibilities** consistent with the observations.

# Perspectives

**In terms of operational application**, this approach is more flexible and provides several advantages:

→ it can focus on a specific region of interest,
→ it can produce targetted products to meet users' requirements,
→ it may serve as a baseline to compare with the dynamical system.

**In terms of method**, the MCMC sampler opens also **new possibilities**:

→ it provides an efficient way to **solve the problem globally** in space and time, with covariance localization,
→ it can cope with fully **general observation constraint** $p(\mathbf{y}^o|\mathbf{x})$ (nonlinear, non-Gaussian, nonlocal),
→ it can be used in 3D as a **possible substitute** to the analysis step in ensemble Kalman filters.

**Difficulties** may come from:

→ the size of the problem and the cost to evaluate the observation constraint,
→ the number of iterations required to reach convergence.

# Complements

# Method : an MCMC sampler based on the Metropolis/Hastings algorithm

**Sample the posterior pdf**
for the evolution of the system
$\mathbf{x}$ ($n \sim 5 \times 10^6$),
given observations $\mathbf{y}^o$ ( $p \sim 10^5$)

$$p(\mathbf{x}|\mathbf{y}^o) \sim p^b(\mathbf{x}) \;\; p(\mathbf{y}^o|\mathbf{x})$$

**Prior ensemble (m=40 members)** → $p^b(\mathbf{x})$

**Observation constraint** → $p(\mathbf{y}^o|\mathbf{x})$

**Anamorphosis transformation**
$\mathbf{x'} = A(\mathbf{x}), \quad \mathbf{x} = A^{-1}(\mathbf{x'})$
to obtain marginally Gaussian $\mathbf{x'}$
(with mean=0 and variance=1):

$$p(\mathbf{x'}|\mathbf{y}^o) \sim p^b(\mathbf{x'}) \;\; p[\, \mathbf{y}^o \,|\, A^{-1}(\mathbf{x'}) \,]$$

**We use local correlations only** → $p^b(\mathbf{x'})$

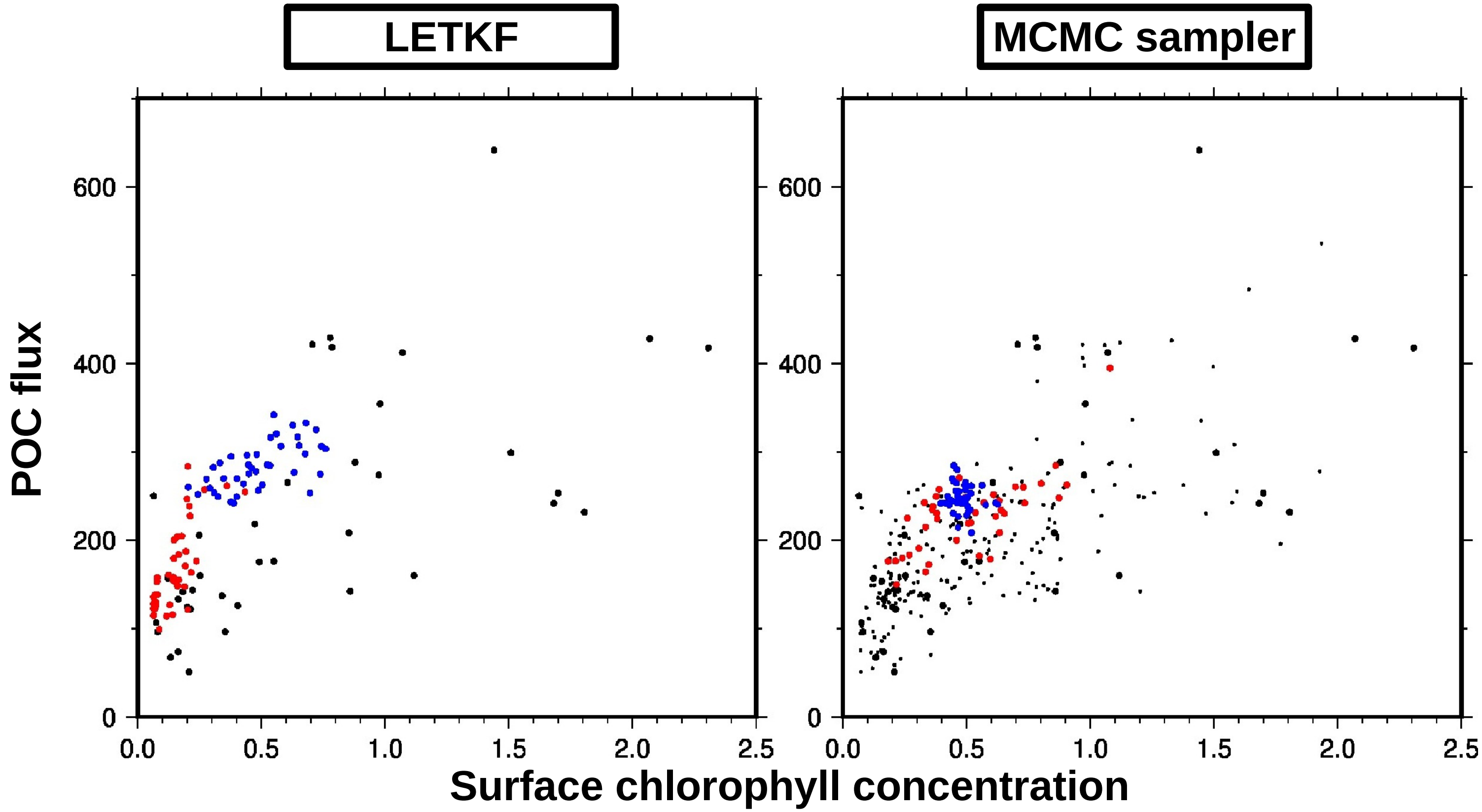
**Kept fully general** → $p[\, \mathbf{y}^o \,|\, A^{-1}(\mathbf{x'}) \,]$

**Iterative method in 2 steps:**

1. **Propose** pseudo-random perturbation of $\mathbf{x'}$ (with cost linear in n)
   → by modulation of an ensemble member with large-scale signals
   ( $\sim 10^{11}$ pseudo-random directions of perturbations)
   → equivalent to a localization of the prior ensemble covariance
2. **Accept/reject** according to cost function: $J^o = -\log p[\, \mathbf{y}^o \,|\, A^{-1}(\mathbf{x'}) \,]$

Brankart J;-M., 2019 : Implicitly Localized MCMC Sampler to Cope With Non-local/Non-linear Data Constraints in Large-Size Inverse Problems. Front. Appl. Math. Stat. 5:58.

# Downward flux of POC at 100 m depth




**Scatterplot between observed quantity and indicator at the center of the region of interest 36°W 47.25°N on May 26**

**In black: prior ensemble simulations from NEMO/PISCES**

**In blue: ensemble analysis using all L3 observations**

**In red: ensemble forecast using L3 observations until May 22**