The complexity of climate model drifts

<u>Davide Zanchettin</u>

Angelo Rubino

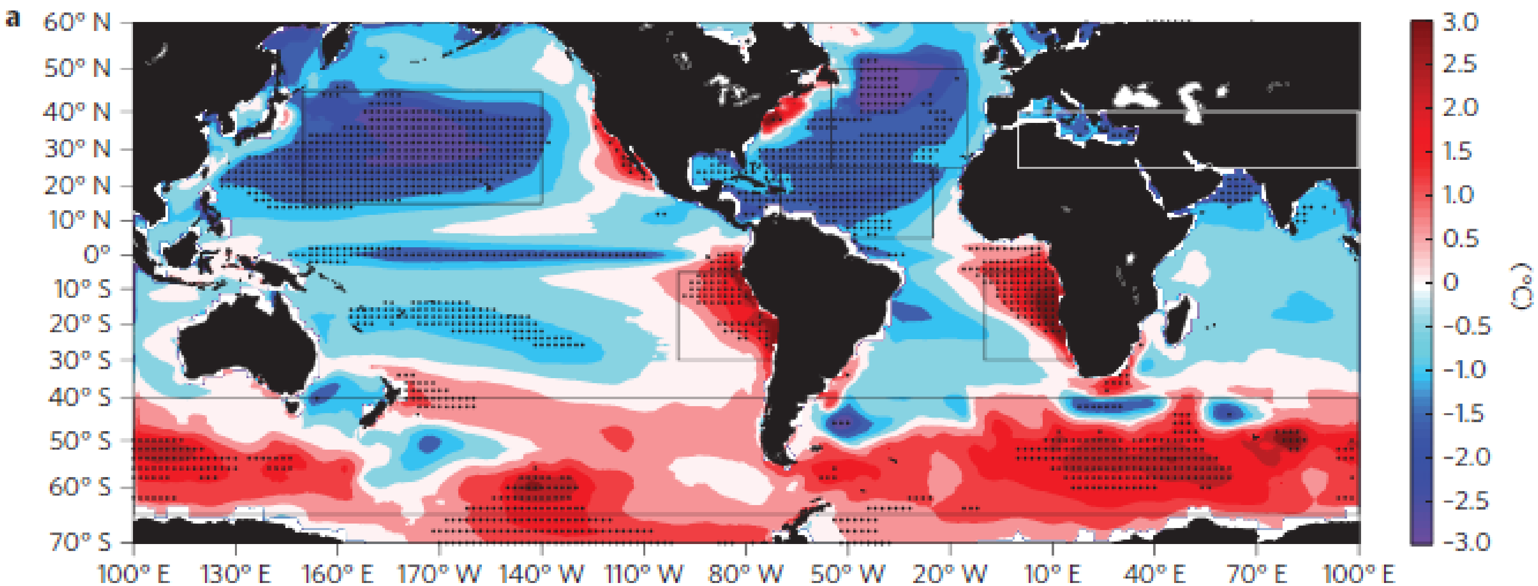
Maeregu Arisido

Carlo Gaetan

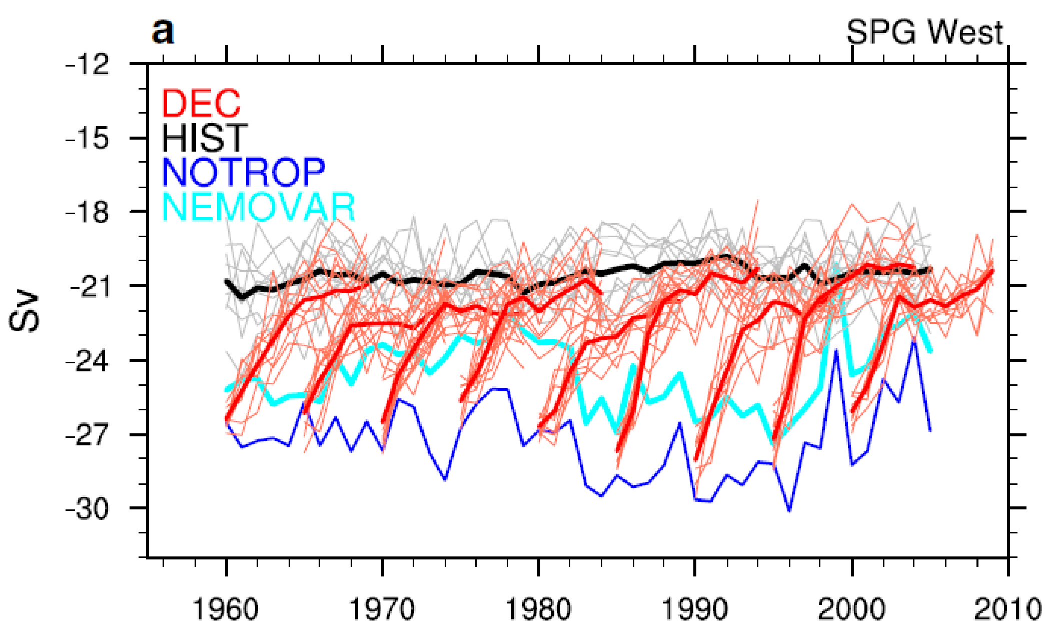
University of Venice, Dept. of Environmetal Sc., Informatics and Statistics

A contribution to PREFACE-WP10: (Statistical methods to assess and improve forecast of Tropical Atlantic variability)

PREFACE-PIRATA-CLIVAR TAV Conference, Paris, UPMC 28-11-206 – 02-12-2016



Wang et al. 2014 | ***Annual-mean SST bias averaged in 22 climate models.*** *The SST bias is calculated by the SST difference between the model SST and extended reconstructed SST.*



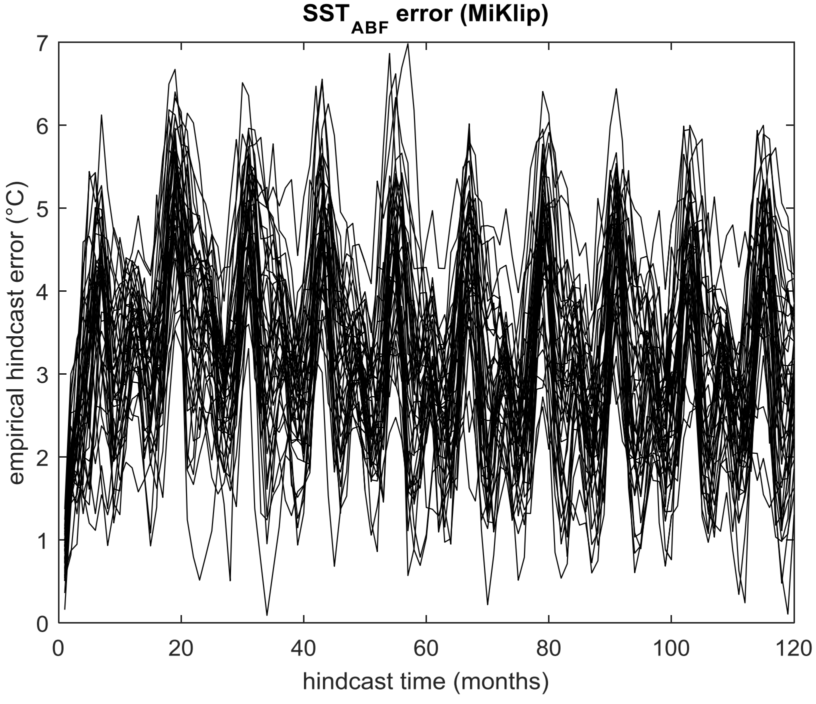
Sanchez-Gomez et al. 2015 | ***Climate model drifts****: Spaghetti plot of the barotropic streamfunction averaged over the western SPG region for decadal hindcasts (DEC, red) and historical simulations (HIST, gray) as a function of leadtime; ensemble means (thick red and black lines).*

Drifts occur at different time scales for different variables, can obscure the initial-condition forecast information and is usually removed a posteriori by an empirical, usually linear, adjustment (IPCC-AR5, 2013)

DCPP guidelines for “data and bias correction for decadal climate predictions”:

$$\overline{Y}_\tau = \frac{1}{n}\sum_{j=1}^{n} Y_{j\tau}$$ forecasts, $j=1,\ldots,n$ initial times; $\tau=1,\ldots,m$ forecast range

$$\overline{X}_\tau = \frac{1}{n}\sum_{j=1}^{n} X_{j\tau}$$ observation-based data



Under full-field initialization $Y_{j\tau=0} \approx X_{j\tau=0}$

the model drift is $\boxed{d_\tau = \overline{Y}_\tau - \overline{X}_\tau}$

and the bias-corrected forecast is:

$$\hat{Y}_{j\tau} = Y_{j\tau} - d_\tau = \overline{X}_\tau + (Y_{j\tau} - \overline{Y}_\tau) = \overline{X}_\tau + Y'_{j\tau}$$

MOTIVATION

We need to better characterize spatial-temporal features of model errors and the uncertainties involved in their estimation and to optimally merge information from observed and simulated data in space and time (it’s the goal of *PREFACE-WP10*).

A STATE-SPACE APPROACH

Dynamical linear models (DLMs) use unobservable state variables which allow direct modelling of the processes generating the observed variability.

$$\boxed{\begin{aligned} y_t &= F\,x_t + v_t, \\ x_t &= G\,x_{t-1} + w_t, \end{aligned}} \quad \begin{aligned} v_t &\sim N(0,V) \\ w_t &\sim N(0,W) \end{aligned}$$

$p(y_t \mid x_t, \boldsymbol{\theta})$ OBSERVATION UNCERTAINTY

$p(x_t \mid x_{t-1}, \boldsymbol{\theta})$ PROCESS UNCERTAINTY

$t = 1,\dots,n$

y_t : *observation vector at time t* $\{p\}$

x_t : *(hidden) state vector at time t* $\{m\}$

G_{mxm} : *system operator*

F_{pxm} : *observation operator*

V_t : *observation error covariance*

W_t : *system error covariance*

θ : *static parameters vector*

BAYESIAN ANALYSIS

$$P(x,\theta \mid y) \propto P(y \mid x,\theta) \cdot P(x \mid \theta) \cdot P(\theta)$$

The DLM formulation can be seen as ***a special case of a general hierarchical statistical model*** *with three levels: data y_t, process x_t, parameters* $\boldsymbol{\theta} = \{G,F,V,W\}$ *(e.g., Cressie and Winkler, 2011).*

The classical ***Kalman filter formulas*** *and* ***Monte Carlo Markov Chain*** *(MCMC) provide efficient and well founded computational tools to determine all the relevant statistical distributions.*

STRUCTURAL DECOMPOSITION OF THE ERROR

The process of interest incorporates systematic contributions to the decadal climate prediction errors:

systematic mean error $\delta(t)$ with stochastic trend $\tau(t)$

annual and semi-annual seasonal biases, namely $\beta^{12}(t)$ and $\beta^{6}(t)$

$$\Delta(t) = \delta(t) + \beta^{12}(t) + \beta^{6}(t)$$

$$\delta(t) = \delta(t\text{-}1) + \tau(t\text{-}1) + \varepsilon_\delta(t) \qquad \varepsilon_\delta \sim N(0, \sigma^2{}_\delta)$$

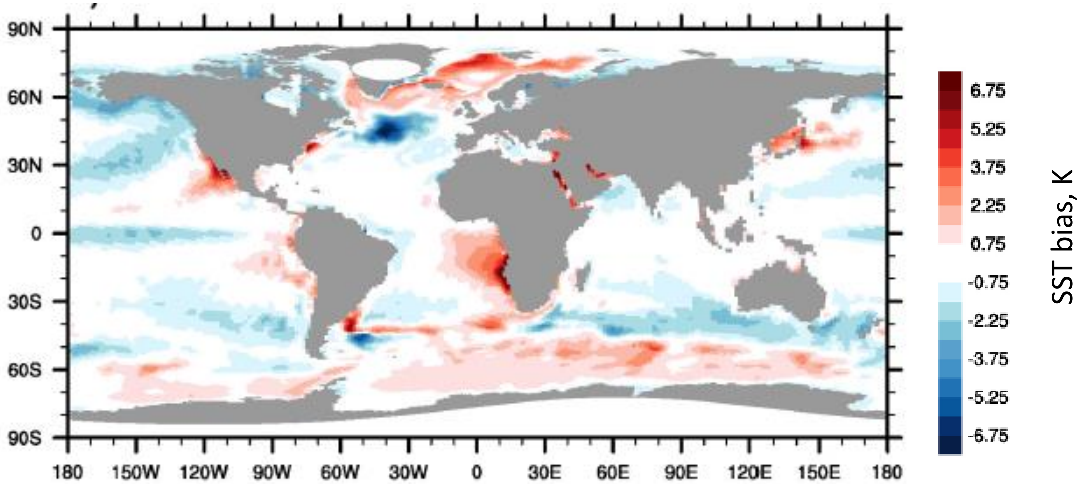
$$\tau(t) = \tau(t\text{-}1) + \varepsilon_\tau(t) \qquad \varepsilon_\tau \sim N(0, \sigma^2{}_\tau)$$

The process model above can be easily extended to include the effect of external factors, by including additional explanatory variables. For one covariate X(t), the model becomes

$$\Delta*(t) = \Delta(t) + \gamma(t)X(t)$$

$$\gamma(t) = \gamma(t\text{-}1) + \varepsilon_\gamma(t) \qquad \varepsilon_\gamma \sim N(0, \sigma^2{}_\gamma)$$

A FIRST APPLICATION of the DLM



From: Jungclaus et al., 2013

Tropical and South Atlantic *monthly*

sea-surface temperatures

from the MiKlip full-field **GECCO 1960-2000 "r1"** decadal hindcasts with MPI-ESM-LR*.

*Bayesian analysis applied on error covariances V and W (a total of **3 parameters**), use **lognormal priors** [logN(0,1)]*

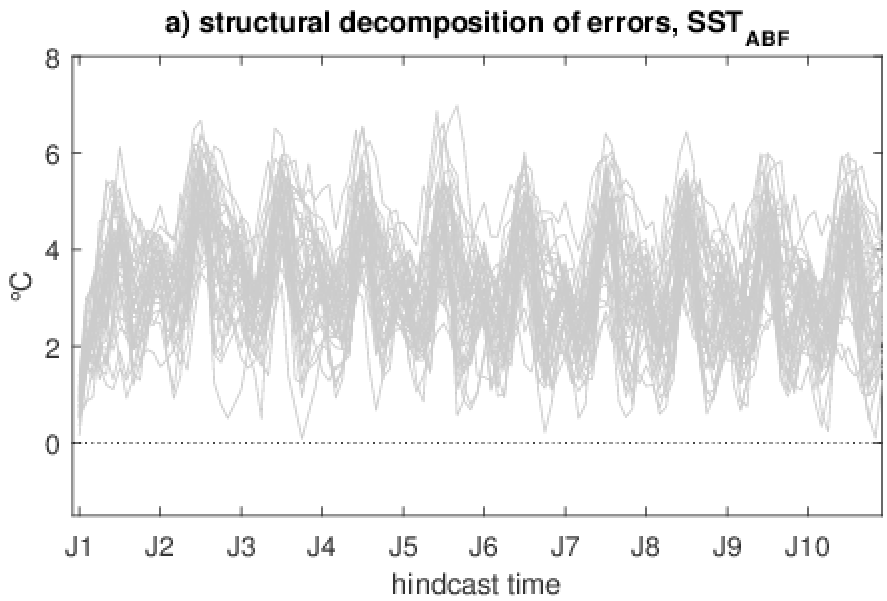
*For spatial analysis, individual **grid points are processed individually**, parallelization speeds up calculation.*

*The MCMC (10000x) is based on the **slicesampler** algorithm.*

*Use the dlmsmo routine from the **dlm toolbox** by Markko Laine*

** Courtesy of Wolfgang Mueller, MPIM*

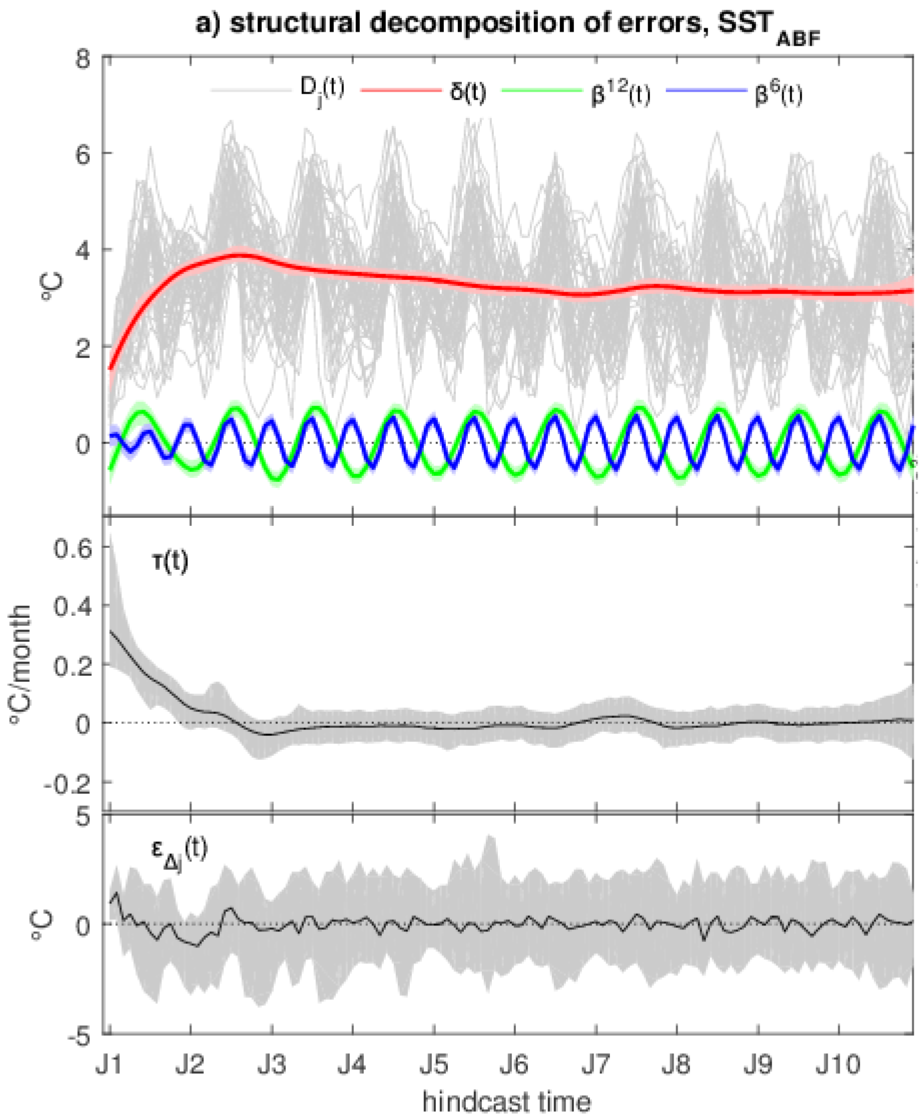
RESULTS – REGIONAL ERROR: SST in the Angola Benguela front



Dj: empirical hindcast error *δ: drift/bias*

τ: stochastic trend component *β: seasonal bias component (annual and semiannual)*

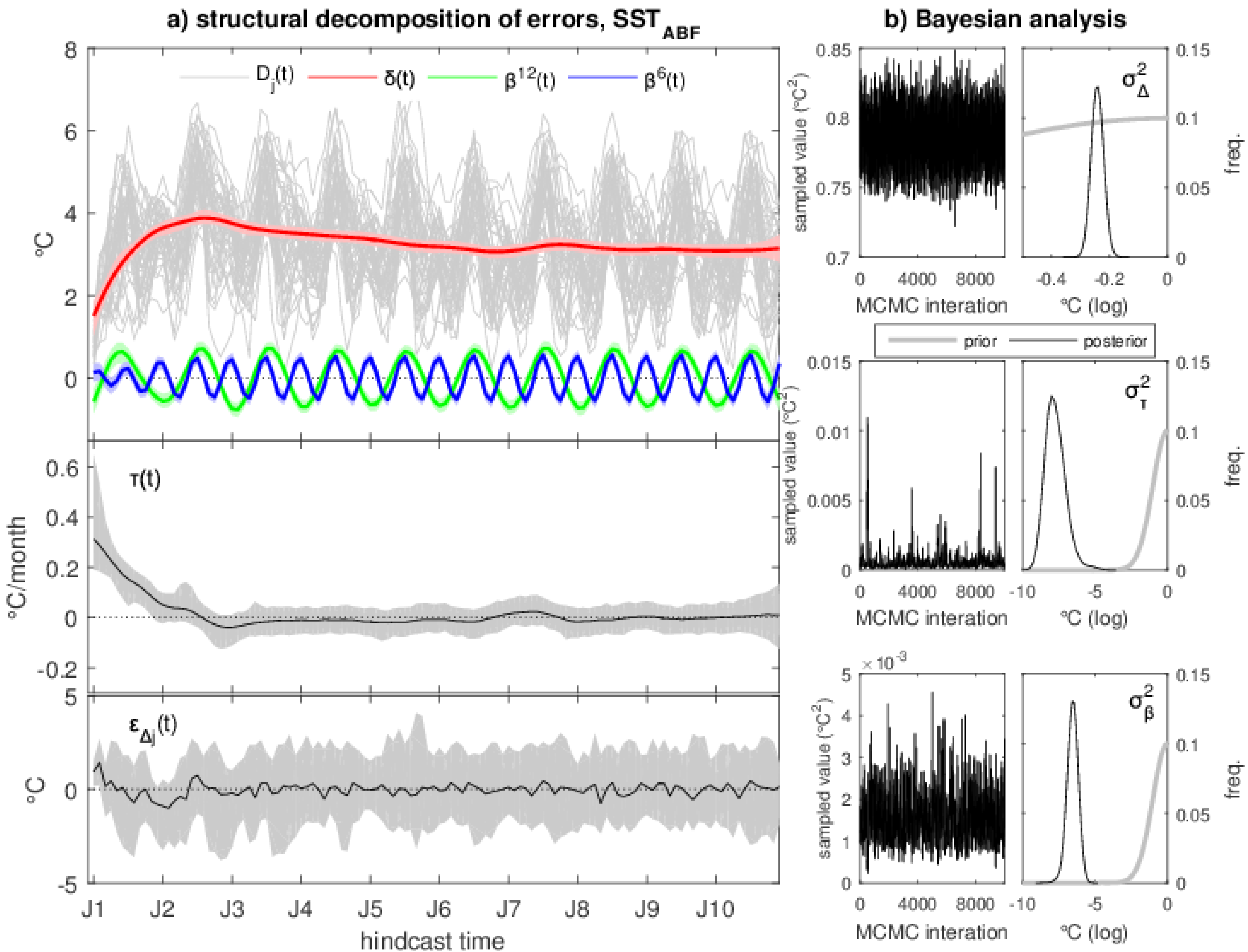
RESULTS – REGIONAL ERROR: SST in the Angola Benguela front



Dj: empirical hindcast error — *δ: drift/bias*

τ: stochastic trend component — *β: seasonal bias component (annual and semiannual)*

RESULTS – REGIONAL ERROR: SST in the Angola Benguela front

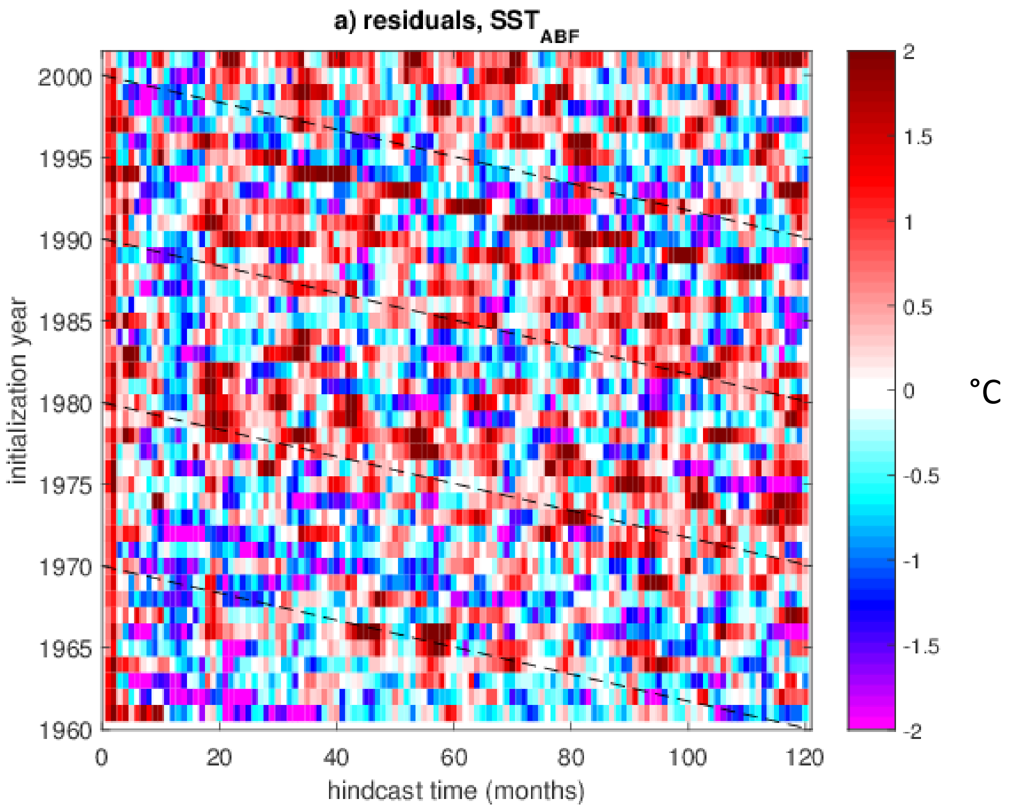


Dj: empirical hindcast error *δ: drift/bias*

τ: stochastic trend component *β: seasonal bias component (annual and semiannual)*

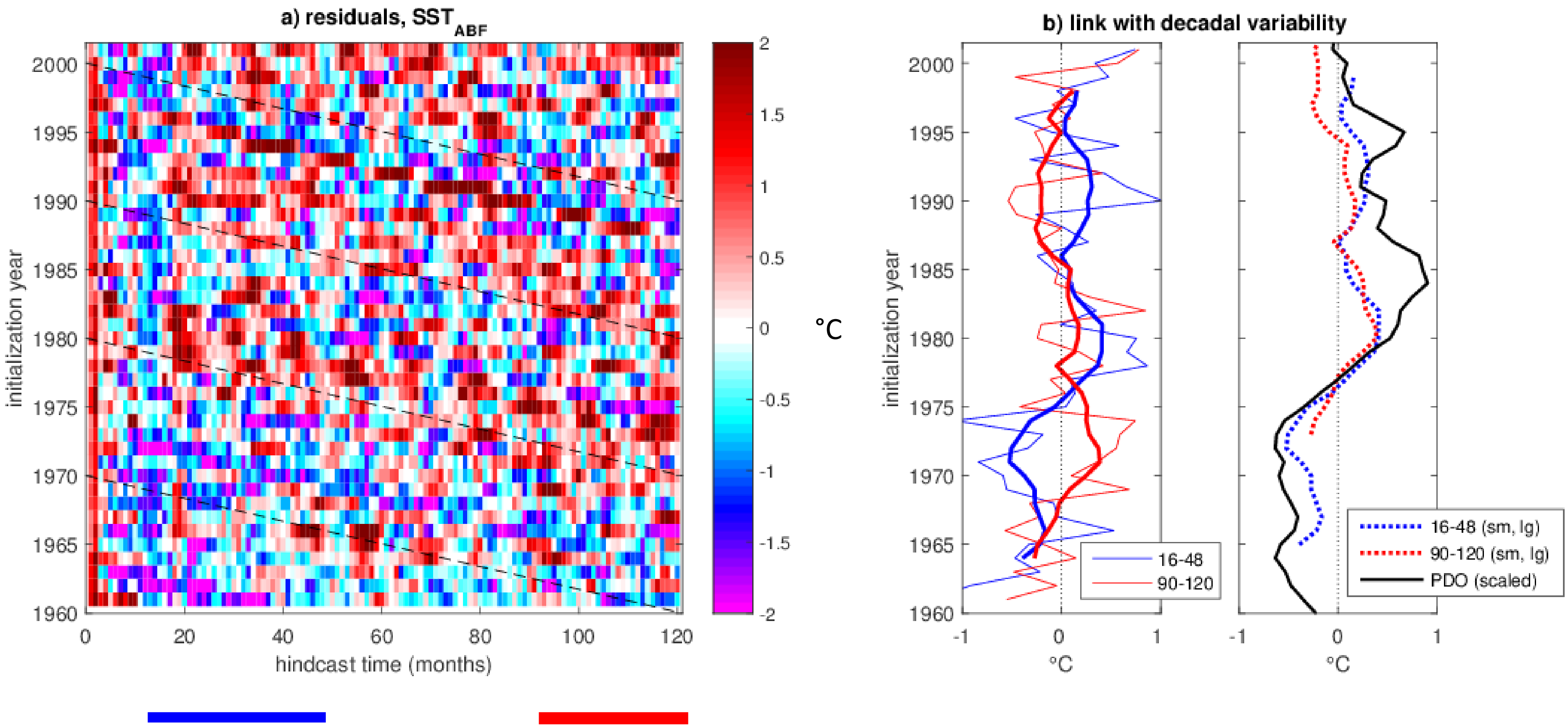
RESULTS – A LOOK AT RESIDUALS (DRIFT-CORRECTED ERRORS), SST_{ABF}

Temporal evolution of posteriori means of monthly-mean residuals in SSTs for the Angola-Benguela front region.



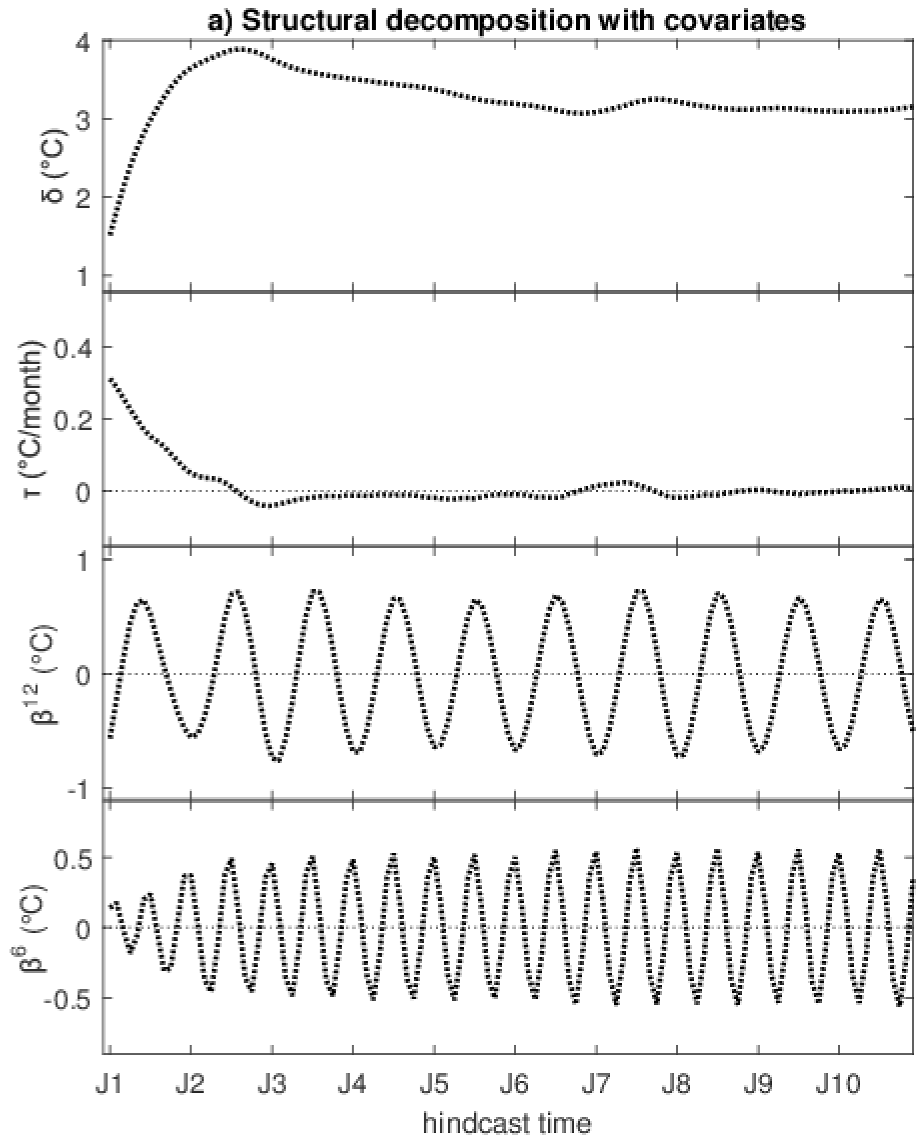
RESULTS – A LOOK AT RESIDUALS (DRIFT-CORRECTED ERRORS), SST_{ABF}

Temporal evolution of posteriori means of monthly-mean residuals in SSTs for the Angola-Benguela front region.



RESULTS – REGIONAL ERROR: SST in the Angola Benguela front

Effect of covariates

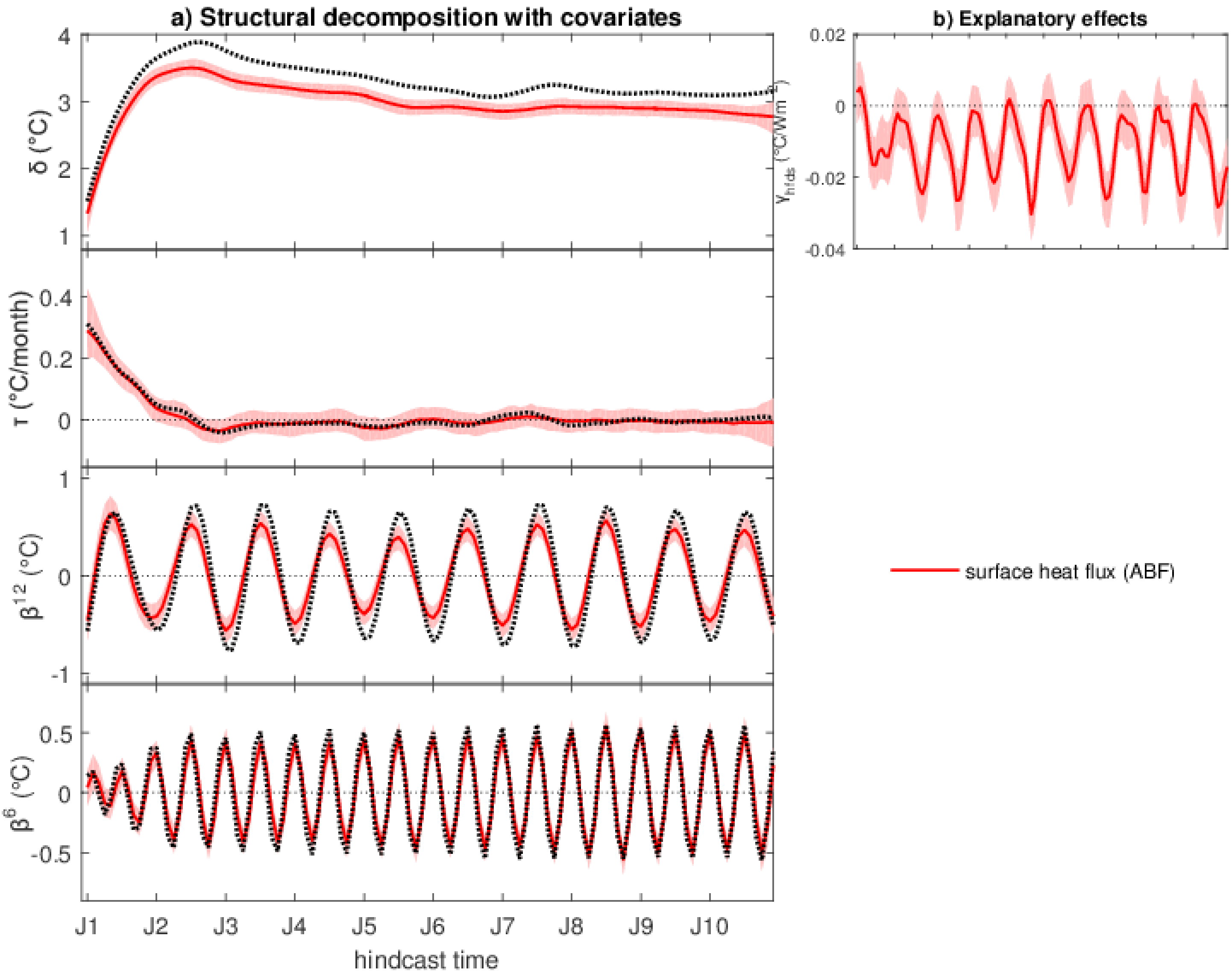


Dj: empirical hindcast error *δ: drift/bias*

τ: stochastic trend component *β: seasonal bias component (annual and semiannual)*

RESULTS – REGIONAL ERROR: SST in the Angola Benguela front

Effect of covariates

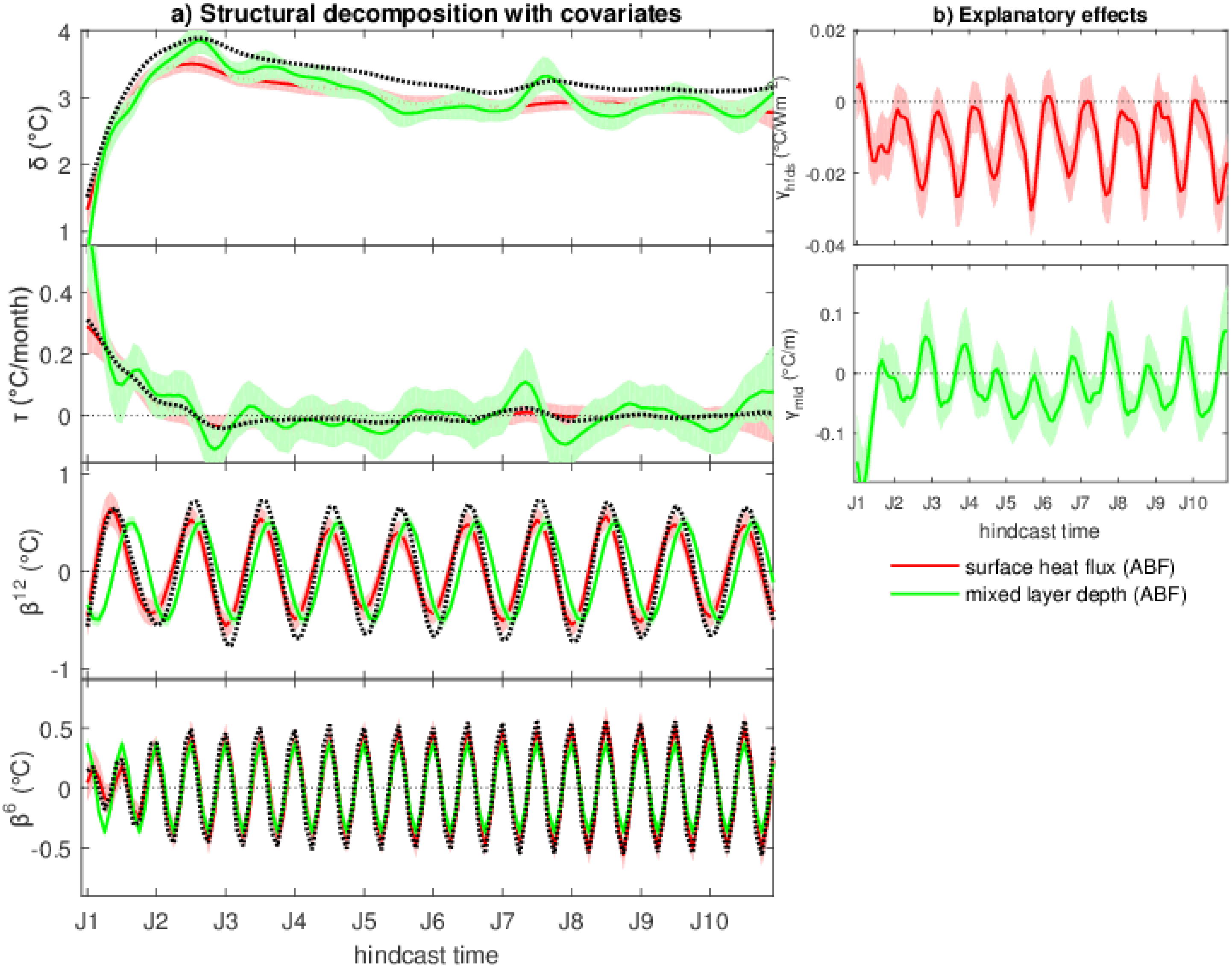


Dj: empirical hindcast error *δ: drift/bias*

τ: stochastic trend component *β: seasonal bias component (annual and semiannual)*

RESULTS – REGIONAL ERROR: SST in the Angola Benguela front

Effect of covariates

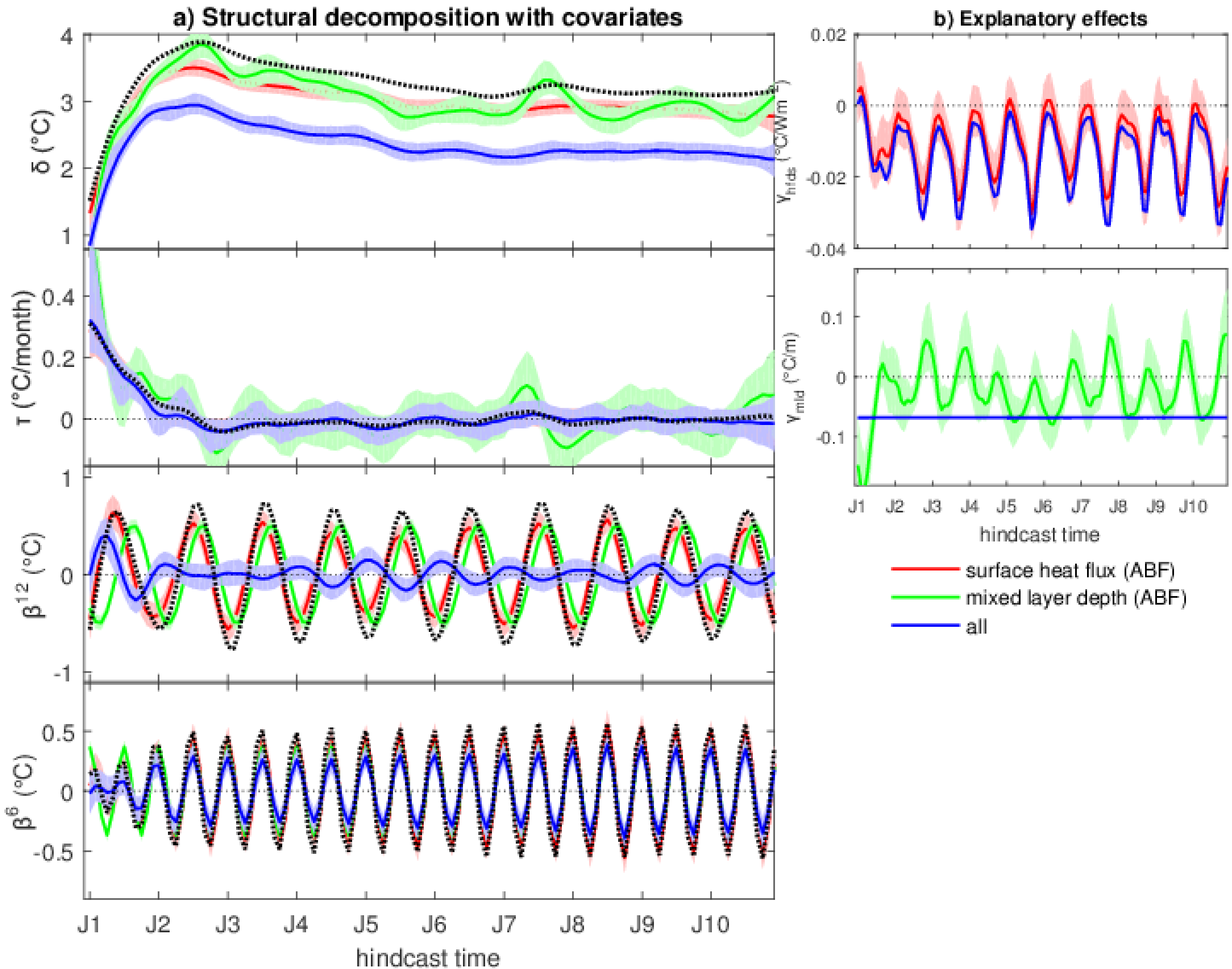


Dj: empirical hindcast error *δ: drift/bias*

τ: stochastic trend component *β: seasonal bias component (annual and semiannual)*

RESULTS – REGIONAL ERROR: SST in the Angola Benguela front

Effect of covariates

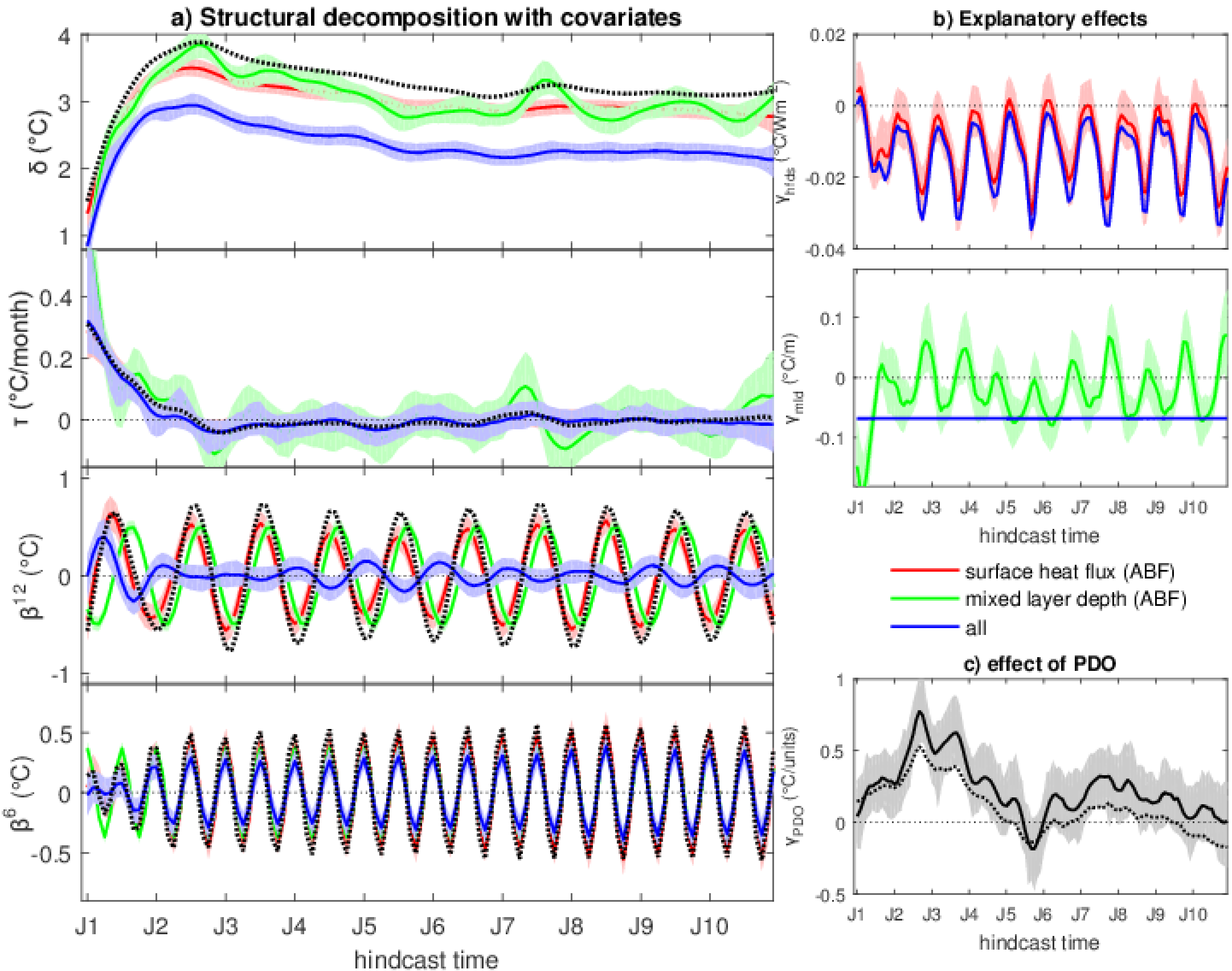


Dj: empirical hindcast error *δ: drift/bias*

τ: stochastic trend component *β: seasonal bias component (annual and semiannual)*

RESULTS – REGIONAL ERROR: SST in the Angola Benguela front

Effect of covariates

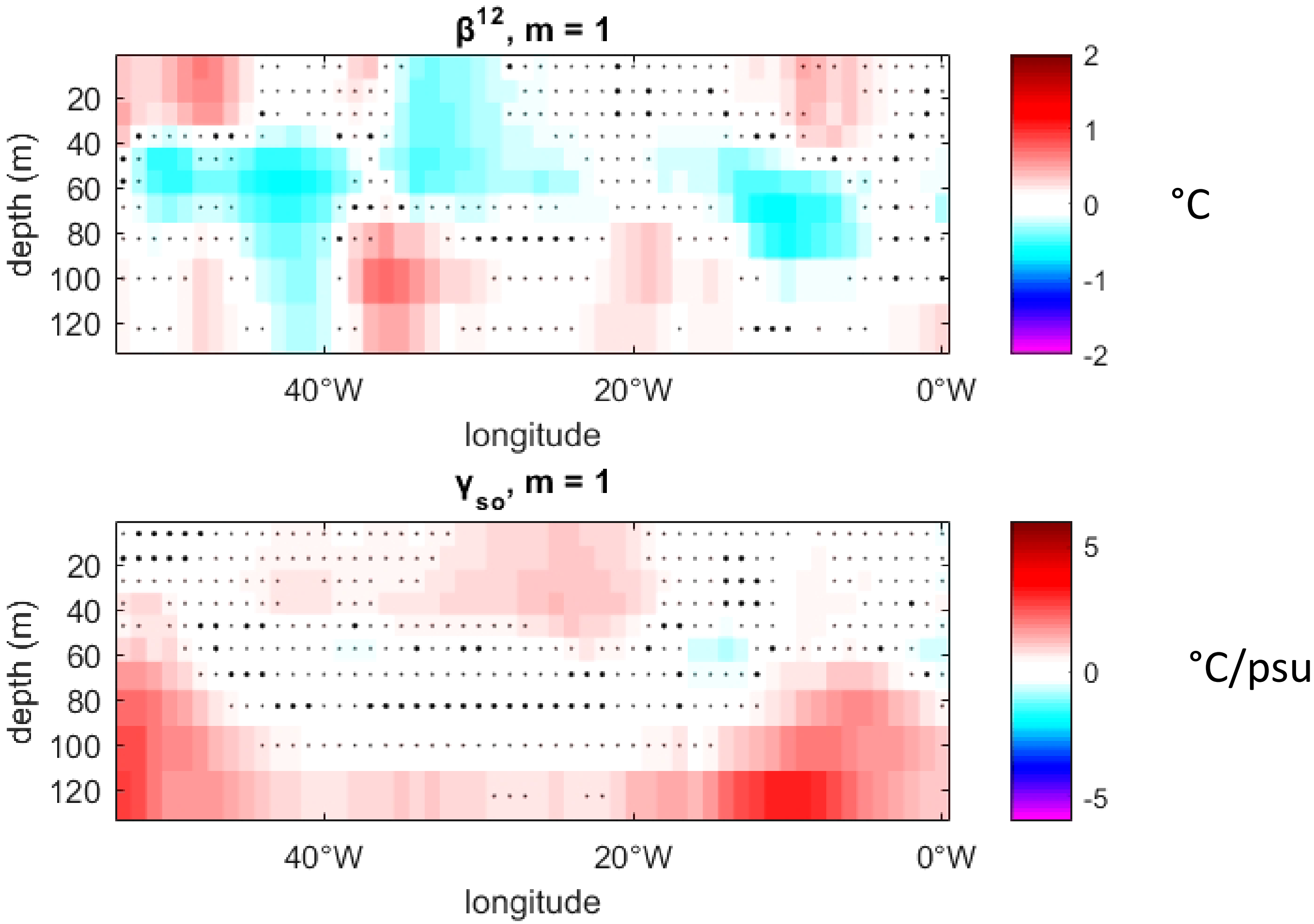


Dj: empirical hindcast error *δ: drift/bias*

τ: stochastic trend component *β: seasonal bias component (annual and semiannual)*

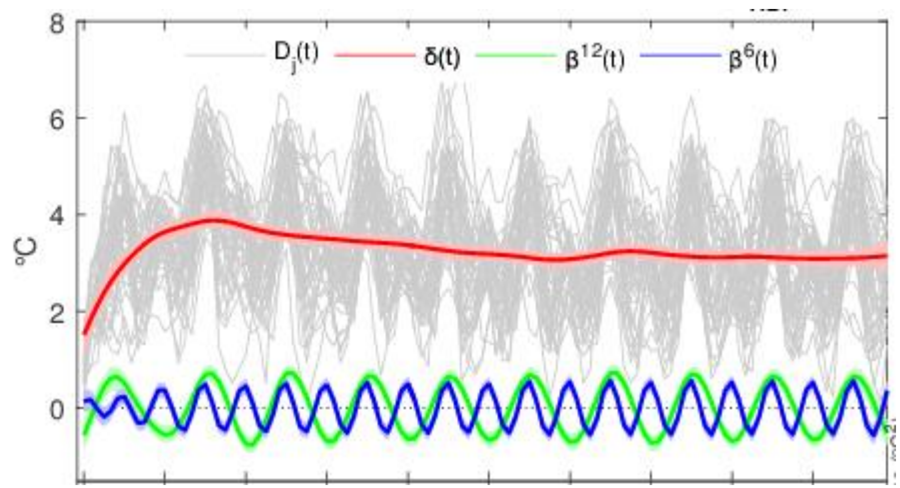
RESULTS – PROPAGATION OF SEASONAL SST ERRORS, THE ROLE OF SALINITY ERRORS

Longitudinal section at 44°S



CONCLUSIONS (WIP) AND OUTLOOK

We propose a **structural decomposition** of **systematic decadal climate prediction errors** (drift/climatological bias and seasonal biases), which is implemented via a state-space model built within a Bayesian hierarchical framework.

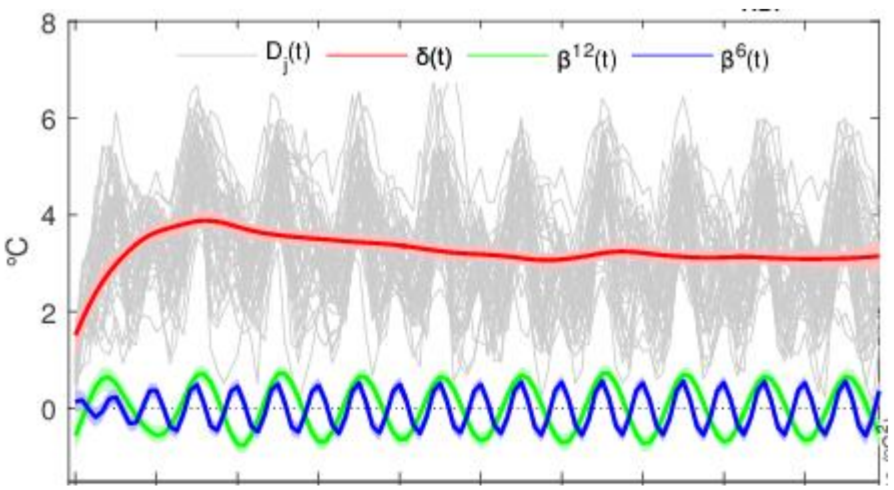


Results help characterizing the **great complexity** behind drift/climatological bias and seasonal biases.
Do we understand the different physical sources, propagation mechanisms and implications of such model error components?

There is an **intimate connection between (estimated) drift development and interdecadal climate evolution**. Furthermore, the hindcast error in a certain location can be substantially shaped by the effect of systematic errors over remote regions (e.g., PDO).
Do the found uncertainties in drift components call for improved drift estimation and adjustment techniques?

CONCLUSIONS (WIP) AND OUTLOOK

We propose a **structural decomposition** of **systematic decadal climate prediction errors** (drift/climatological bias and seasonal biases), which is implemented via a state-space model built within a Bayesian hierarchical framework.



Results help characterizing the **great complexity** behind drift/climatological bias and seasonal biases.

Do we understand the different physical sources, propagation mechanisms and implications of such model error components?

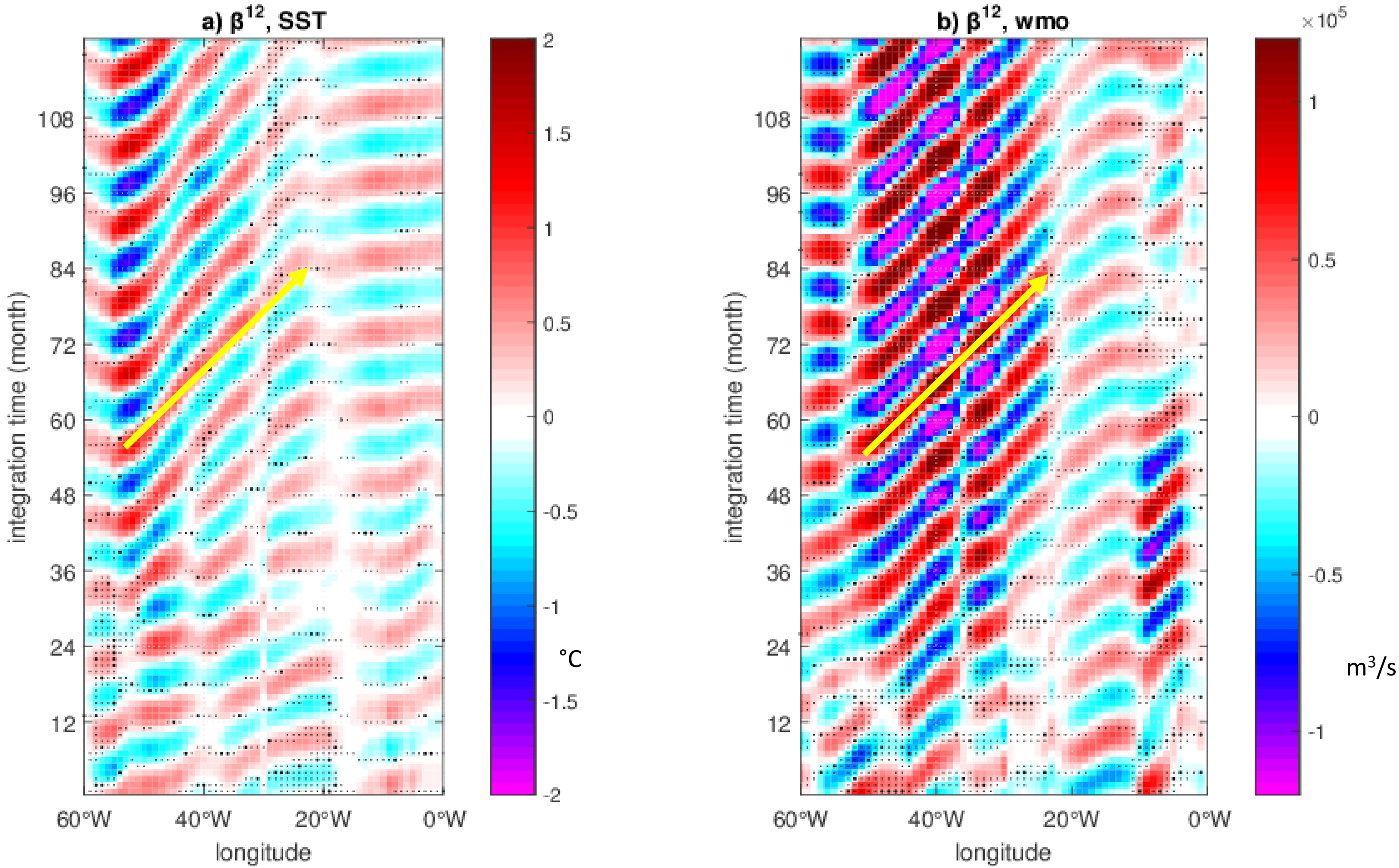
There is an **intimate connection between (estimated) drift development and interdecadal climate evolution**. Furthermore, the hindcast error in a certain location can be substantially shaped by the effect of systematic errors over remote regions (e.g., PDO).

Do the found uncertainties in drift components call for improved drift estimation and adjustment techniques?

THANK YOU FOR YOUR ATTENTION

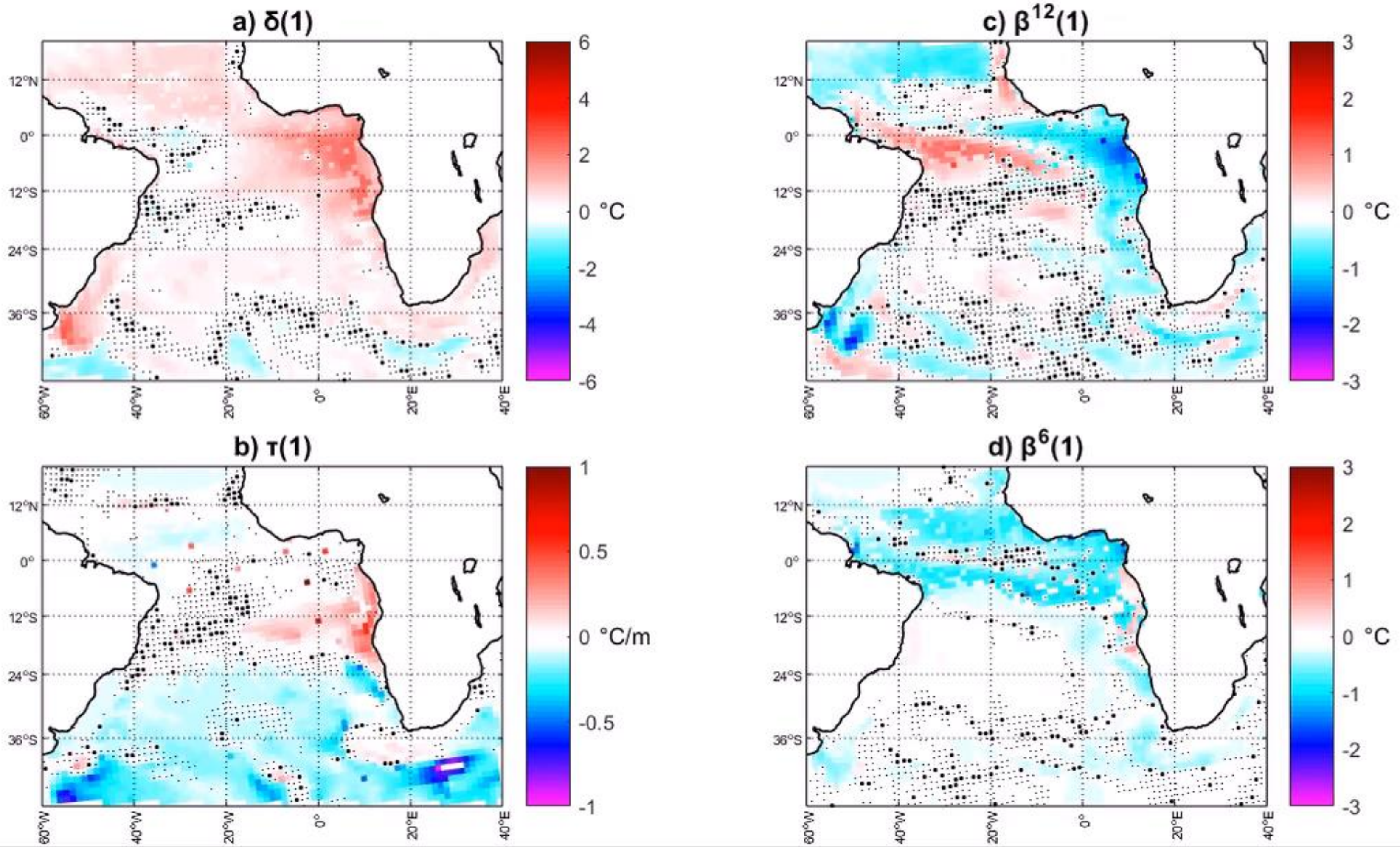
RESULTS – PROPAGATION OF SEASONAL SST ERRORS

→ Pulse error signals generated around 50°W apparently travel eastward to about 25°W, with a speed of approximately 4 cm/s



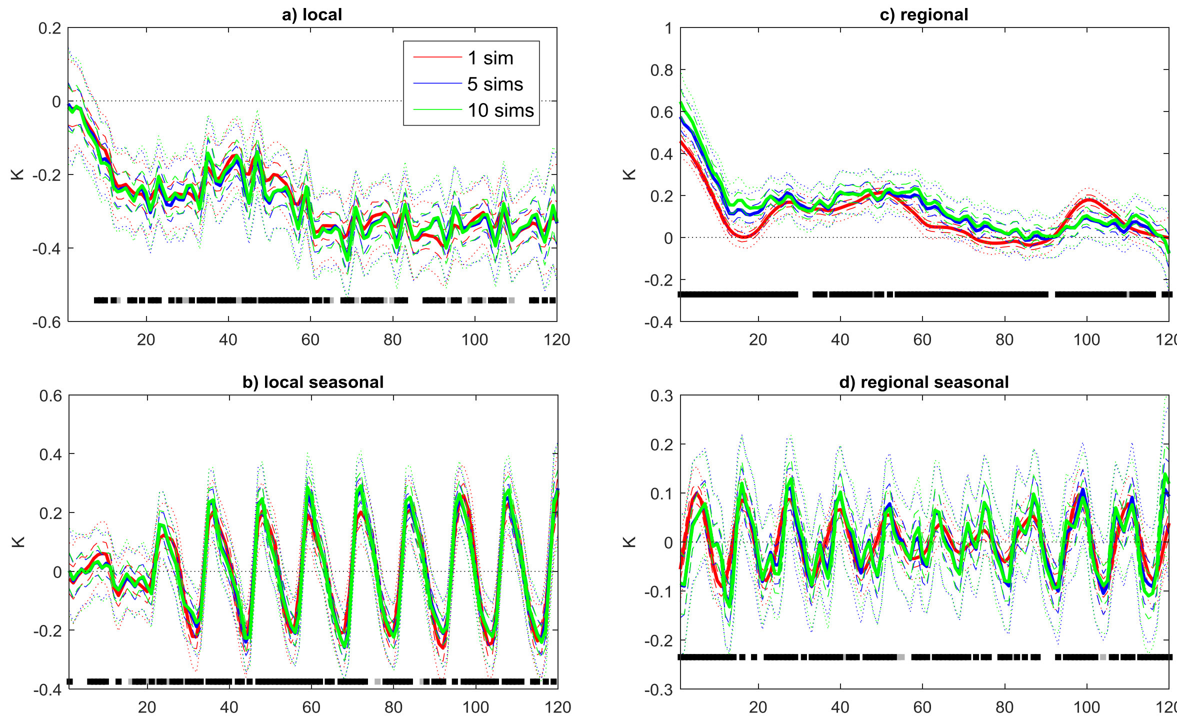
RESULTS - MARGINAL POSTERIOR DISTRIBUTIONS OF SST ERRORS

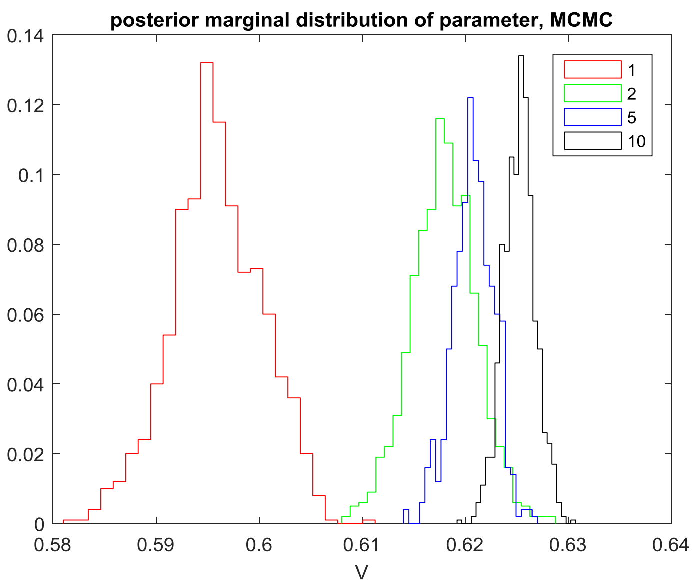
grid-point analysis

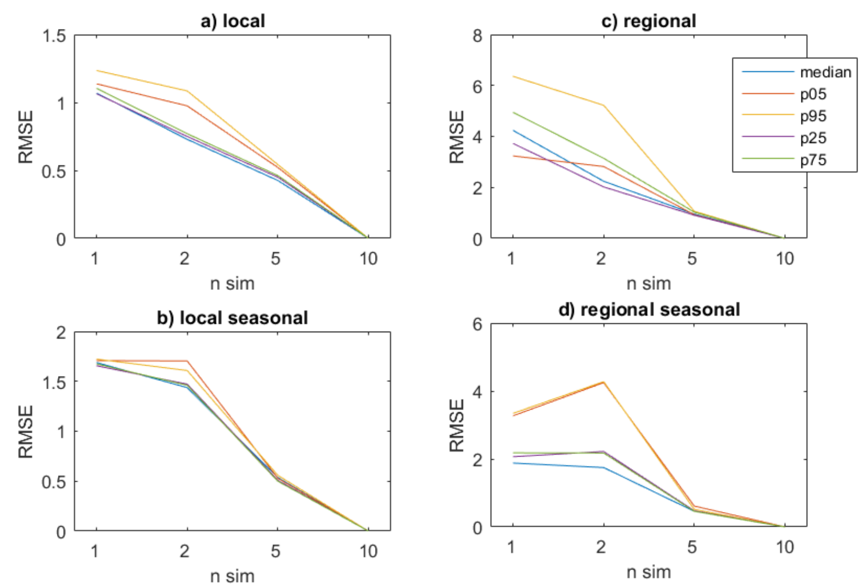


Shading: posteriori median (drift component)*; large (small) dots mark grid points where the 0-value lies within the 40th-60th (5th-95th) percentile range of the posteriori distribution*

RESULTS – IMPACTS OF NUMBER OF “OBSERVATIONS” ON DRIFT ESTIMATION







A STATE-SPACE APPROACH

We use DLM to determine:

uncertainty of unknown states and their evolution conditional to observations and model parameters:

$$p(x_{1:n} \mid y_{1:n}, \theta)$$

by means of Kalman based simulation smoother

Uncertainty of unknown states and parameters and their evolution conditional to all available observations (Bayesian approach):

$$p(x_{1:n} \mid y_{1:n}) = \int p(x_{1:n}, \theta \mid y_{1:n})\, d\theta$$

by means of Monte Carlo Markov Chain (MCMC). This is possible thanks to the Markov property inherent in the definition of our model: the state at time t is statistically conditionally independent on the whole history as it only depends on the state at t-1.

1. KF forward recursion

Assuming the initial distributions at time t=0 are known, the Kalman filter forward recursion can be used to calculate the distribution of the state vector x_t, given observations up to time t: $p(x_t \mid y_t, \theta)$.

This is done by calculating, as **prior**, the mean and covariance matrix of one-step-ahead predicted states: $p(x_t \mid x_{t-1}, y_{t-1}, \theta) = N(x^-_t, C^-_t)$

$$\hat{x}_t = G_t\, \bar{x}_{t-1} \qquad prior\ mean\ for\ x_t$$

$$\hat{C}_t = G_t\, \bar{C}_{t-1}\, G_t^T + W_t \qquad prior\ covariance\ for\ x_t$$

$$C_{y,t} = F_t\, \hat{C}_{t-1}\, F_t^T + V_t \qquad covariance\ for\ predicting\ y_t$$

Then the **posterior** state and its covariance are calculated using the Kalman gain matrix:

$$K_t = \hat{C}_t\, F_t^T C_{y,t}^{-1} \qquad Kalman\ gain$$

$$v_t = y_t - F_t\, \hat{x}_t \qquad prediction\ residual$$

$$\bar{x}_t = \hat{x}_t + K_t\, v_t \qquad posterior\ mean\ for\ x_t$$

$$\bar{C}_t = \hat{C}_t - K_t F_t \hat{C}_t \quad posterior\ covariance\ for\ x_t$$

Equations are iterated for t=1,…,N

2. Kalman smoother backward recursion

KF provides distributions of xt given observations up to time t. We want to **account for all observations, so:** $p(x_t|y_{1:n},\theta)$ (all gaussian). The Kalman smoother backward recursion provide so-called smoothed states for t=N,N-1,…,1. Setting r_{N+1} and N_{N+1} equal to zero:

$$L_t = G_t - G_t K_t F_t \quad auxiliary\ variable$$

$$r_t = F_t^T C_{y,t}^{-1} v_t + L_t^T r_{t+1} \quad auxiliary\ variable$$

$$N_t = F_t^T C_{y,t}^{-1} F_t + L_t^T N_{t+1} L_t \quad auxiliary\ variable$$

$$\tilde{x}_t = \hat{x}_t + \hat{C}_t r_t \quad smoothed\ state\ mean$$

$$\tilde{C}_t = \hat{C}_t - \hat{C}_t N_t \hat{C}_t \quad smoothed\ state\ covariance$$

3. We need full joint posteriori distribution of all states given all observations (see 1 and 2) and parameters: $p(x_{1:N}\ |\ y_{1:N},\ \theta)$. This distribution does not have a closed form solution, but we **can draw** realizations for it using the so-called simulation smoother algorithm.

In practice, the algorithm proceeds as follows:

- Sample from state space equations to get $x'_{1:N}$ and $y'_{1:N}$ (' stands for tilde, smoothed values)

- Use Kalman smoother with the new observation $y'_{1:N}$ to get smoothed states $x'^{s}_{1:N}$

- Add the state residual to the original smoothed states to obtain $x^*_{1:N} = x'_{1:N} - x'^{s}_{1:N} + x^{s}_{1:N}$

4. Uncertainty on parameters

We do not want θ to be fixed, instead we want to estimate it using Bayesian statistics. We need the marginal likelihood function $p(y_{1:n} \mid \theta)$ with the uncertainty of states accounted for (which means integrated out). For each θ, such likelihood is provided as a byproduct of the Kalman filter.

Due to the Markov property of the state space equations, we can calculate the marginal likelihood as:

$$p(y_{1:N} \mid \theta) = p(y_1 \mid \theta) \prod_{t=2}^{N} p(y_t \mid y_{1:t-1}, \theta)$$

Which for a Gaussian linear model is proportional to:

$$\propto \exp\left\{ -\frac{1}{2} \sum_{t=1}^{N} \left[(y_t - F_t \hat{x}_t)^T C_{y,t}^{-1} (y_t - F_t \hat{x}_t) + \log(|C_{y,t}|) \right] \right\}$$

5. MCMC

A MCMC is performed to calculate the marginal posterior distribution $p(\theta \mid y_{1:N})$, using the likelihood defined in step 4 and with proper priors.

6. Steps 5 and 1-3 are combined to draw samples from the distribution $p(x_{1:N}, \theta \mid y_{1:N})$

BAYESIAN ANALYSIS

We can apply the Bayesian inference on error covariances W and V. We must specify priors (all Gaussian) and likelihoods for all such unknown parameters.

Practically, to reduce computational requirements, we define priors/likelihoods for the standard deviations of the following parameters:

one prior for V (actually fixed and not estimated in present analysis of MiKlip hindcasts)

four priors for W (one for DF, one for B, one common for SF1 and SF2, one common for BSF1 and BSF2).

An adaptive Metropolis algorithm is iteratively used to sample from the full posterior distribution of the unknown parameters.

Kalman filter and Kalman smoother are then used to iteratively sample the system states along the MCMC (i.e., we derive associated marginal distributions for each of the state components)