Метод эмпирического анализа дальних связей в климатической системе Земли

Е.М. Лоскутов, П.А. Мурзина, Д.Н. Мухин, С.Е. Сафонов

Институт прикладной физики им. А.В. Гапонова-Грехова РАН

The method for empirical analysis of teleconnections in the Earth's climate system

E.M. Loskutov, P.A. Murzina, D.N. Mukhin, S.E. Safonov

A.V. Gaponov-Grekhov Institute of Applied Physics, Russian Academy of Sciences

Исследование выполнено при поддержке РНФ (проект № 23-62-10043, https://rsci.ru/project/23-62-10043/)

Presentation outline

- **Problem statement, choice of modeling area**
- **Empirical analysis: construction of data-driven (empirical) models**
  - **The evolution operator form, Bayesian approach**
- **The method of considering interactions between basins**
- **Results**
  - **Comparison of methods for accounting for connectivity**
  - **Improved forecast due to accounting for connectivity**
- **Conclusions**

Problem statement, choice of modeling area

Goal:

Empirical study of the dynamics of climatic characteristics (in particular, SST) in different basins of the world ocean in order to improve their predictability.

Pacific ocean

Objective

Revealing of interbasin teleconnections by data-driven (empirical) method. Comparison of the ESMs with the reanalysis data and with each other.

DATA:

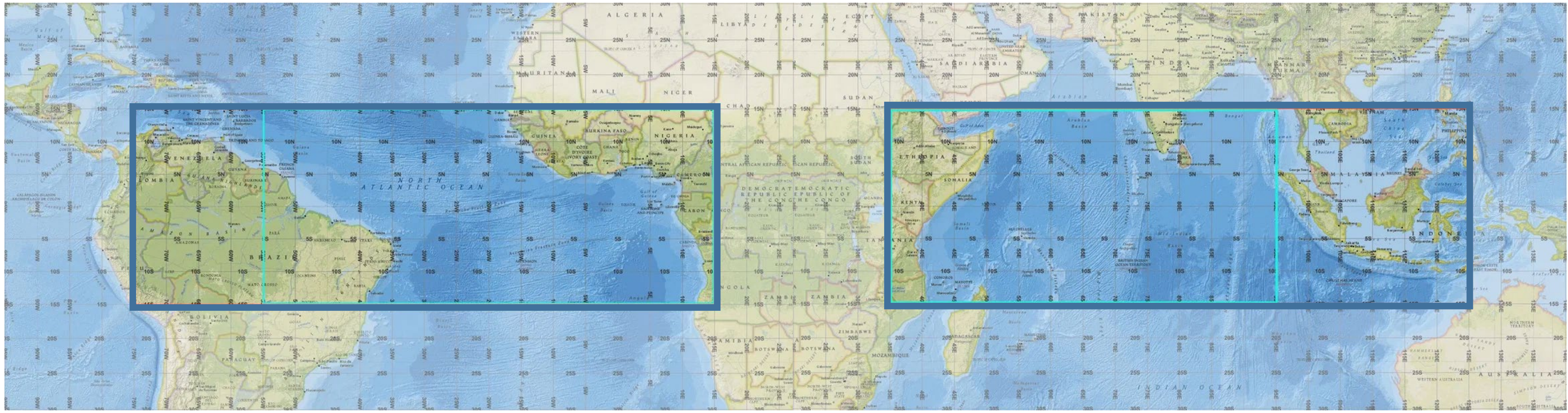
Monthly sea surface temperature anomalies in the tropical belt

from 1960 to 2021 years for **ERSST** data and **INMCM** model;

from 1960 to 2014 years for **Hadley** and **MPI** models

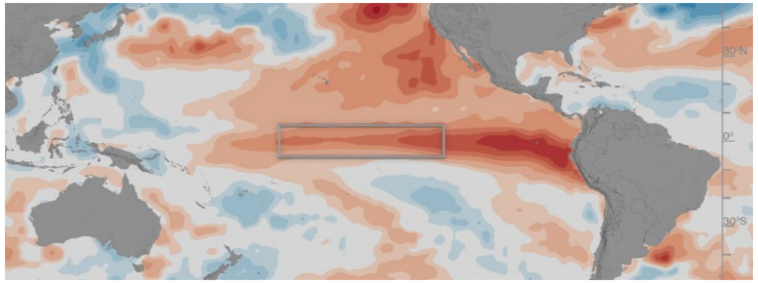
Atlantic ocean

Indian ocean



Empirical analysis: construction of data-driven (empirical) models

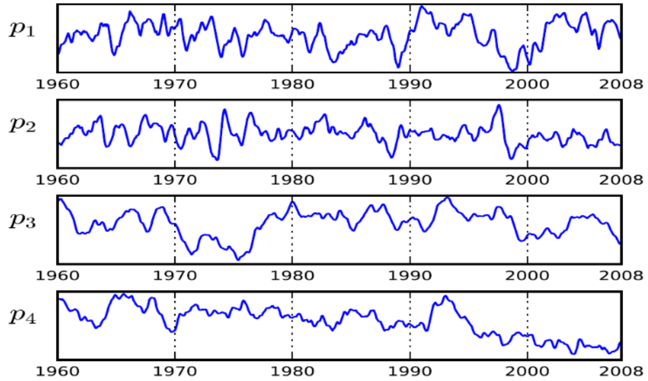
Dataset (ERSST or ESM (INMCM, Hadley, MPI)) $\mathbf{x}_1, \dots, \mathbf{x}_N$



$\mathbf{x}_n \in \mathbb{R}^D$, D is ***very high*** (~number of grid nodes)

Mapping onto principal manifold

$\mathbf{x}_1, \dots, \mathbf{x}_N \rightarrow \mathbf{p}_1, \dots, \mathbf{p}_N$, $\mathbf{p}_n \in \mathbb{R}^d, d \ll D$



data-driven model

Evolution operator on the manifold

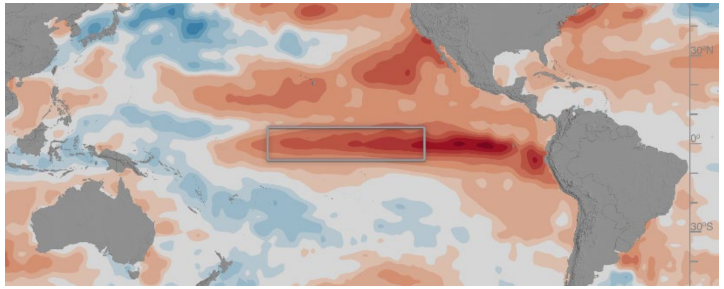
$$\mathbf{p}_{n-L}, \dots, \mathbf{p}_{n-1} \rightarrow \mathbf{p}_n$$

(optimal, nonlinear, non-stationary, stochastic)

Prediction of principal components

$$\mathbf{p}_1, \dots, \mathbf{p}_N \rightarrow \mathbf{p}_{N+1}, \mathbf{p}_{N+2}, \dots$$

Prediction of observed variables (SST)



$$\mathbf{p}_{N+1}, \mathbf{p}_{N+2}, \dots \rightarrow \mathbf{x}_{N+1}, \mathbf{x}_{N+2}, \dots$$

Empirical analysis: construction of data-driven (empirical) models

Evolution operator

$$\mathbf{X}_{n-L}, \dots, \mathbf{X}_{n-1} \to \mathbf{X}_n \quad (\mathbb{R}^{d\times L} \to \mathbb{R}^d)$$

1. Nonlinear
2. Stochastic
3. Forced
4. Non-stationary

external forcing (3)

long-term change (non-stationarity (4))

$$\boxed{\boldsymbol{X}_n = \boldsymbol{f}(\mathbf{X}_{n-1}, \dots, \mathbf{X}_{n-L}; \mathbf{F}_n\ , t_n) + \hat{\mathbf{g}}\ (\mathbf{X}_{n-1}, \dots, \mathbf{X}_{n-L})\boldsymbol{\xi}_n}$$

Stochastic term (2) (unresolved processes – short timescales)
state depended, ξ - delta-correlated

Nonlinear (1) approximator (artificial neural network - ANN)

$$\boldsymbol{f}(\mathbf{X}; \mathbf{F}, t) = \sum_{i=1}^{m} (\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i t) \cdot \tanh(\boldsymbol{\omega}_i \mathbf{X} + \mathbf{v}_i \mathbf{F} + \gamma_i)$$

$$\hat{\mathbf{g}}(\mathbf{X}) = \sum_{i=1}^{m} \hat{\boldsymbol{\alpha}}_i \cdot \tanh(\boldsymbol{\omega}_i \mathbf{X} + \gamma_i)$$

Empirical analysis: construction of data-driven (empirical) models

Model learning and optimization

We need:

- Optimizing the hyperparameters (L, m_f) (here $\mathrm{m}_g = 0$)
- Learning the parameters $\boldsymbol{\mu} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \hat{\mathrm{g}})$

$$\boxed{\mathbf{X}_n = \boldsymbol{f}(\mathbf{X}_{n-1}, \dots, \mathbf{X}_{n-L}; \mathbf{F}_n, t_n) + \hat{\mathbf{g}}\, \boldsymbol{\xi}_n}$$

$$\boldsymbol{f}(\mathbf{X}; \mathbf{F}, t) = \sum_{i=1}^{\mathrm{m}_f} (\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i t) \cdot \tanh(\boldsymbol{\omega}_i \mathbf{X} + \boldsymbol{\nu}_i \mathbf{F} + \gamma_i)$$

$$\boldsymbol{\mu} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \hat{\mathbf{g}})$$

Bayes' theorem

$$P(H_i|\overline{X}) = \frac{P(\overline{X}|H_i) * P(H_i)}{\sum_i P(\overline{X}|H_i) * P(H_i)}$$

here H_i is a model with $(L, \mathrm{m}_f)_i$ hyperparameters,

$P(\overline{X}|H_i)$ is the probability of obtaining the observed dataset $\overline{X}$ from all possible data generated by the model H_i (Bayesian evidence of the model H_i),

$P(H_i)$ - the probability of a model reflecting a priori knowledge about the given model (a prior probability).

The method of considering interaction between basins

$$X_n = \boldsymbol{f}(\mathbf{X}_{n-1}, \dots, \mathbf{X}_{n-L}; \mathbf{F}_n, t_n) + \hat{\mathbf{g}}(\mathbf{X}_{n-1}, \dots, \mathbf{X}_{n-L})\boldsymbol{\xi}_n,$$

$$\boldsymbol{f}(\mathbf{X}; \mathbf{F}, t) = \sum_{i=1}^{m} (\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i t) \cdot \tanh(\boldsymbol{\omega}_i \mathbf{X} + \mathbf{v}_i \mathbf{F} + \gamma_i)$$

A **single** (joint) **model** for all basins. Data from all basins are combined into one dataset.

Several models with forcing. One isolated model per basin. Data from other basins is used as a forcing.

$$\mathbf{X}_n^* = \boldsymbol{f}(\mathbf{X}_{n-1}^*, \dots, \mathbf{X}_{n-L}^*; t_n) + \hat{\mathbf{g}}\,\boldsymbol{\xi}_n$$

$$X_n = \boldsymbol{f}(\mathbf{X}_{n-1}, \dots, \mathbf{X}_{n-L}; \mathbf{F}_n, t_n) + \hat{\mathbf{g}}\,\boldsymbol{\xi}_n$$

$\mathbf{X}^*$ Dataset from all basins

$\mathbf{X}$ Time series from basin of interest

$\mathbf{F}$ Time series from other basins

The method of considering interaction between basins

Isolated models

Several models with forcing. One isolated model per basin.
Data from other basins is used as a forcing.

Forced models:

$$\boldsymbol{F}^{\mathrm{I,P}} = (\boldsymbol{X}^{\mathrm{Indian}}, \boldsymbol{X}^{\mathrm{Pacific}})$$

$$\boldsymbol{F}^{\mathrm{A,P}} = (\boldsymbol{X}^{\mathrm{Atlantic}}, \boldsymbol{X}^{\mathrm{Pacific}})$$

$$\boldsymbol{F}^{\mathrm{A,I}} = (\boldsymbol{X}^{\mathrm{Atlantic}}, \boldsymbol{X}^{\mathrm{Indian}})$$

$$\boldsymbol{X}_n^{\mathrm{A}} = \boldsymbol{f}\left(\boldsymbol{X}_{n-1}^{\mathrm{A}}, \ldots, \boldsymbol{X}_{n-L}^{\mathrm{A}}, \boldsymbol{F}^{\mathrm{I,P}}\right) + \widehat{\boldsymbol{g}}^{\star}\xi_n$$

$$\boldsymbol{X}_n^{\mathrm{I}} = \boldsymbol{f}\left(\boldsymbol{X}_{n-1}^{\mathrm{I}}, \ldots, \boldsymbol{X}_{n-L}^{\mathrm{I}}, \boldsymbol{F}^{\mathrm{A,P}}\right) + \widehat{\boldsymbol{g}}^{\star}\xi_n$$

$$\boldsymbol{X}_n^{\mathrm{P}} = \boldsymbol{f}\left(\boldsymbol{X}_{n-1}^{\mathrm{P}}, \ldots, \boldsymbol{X}_{n-L}^{\mathrm{P}}, \boldsymbol{F}^{\mathrm{A,I}}\right) + \widehat{\boldsymbol{g}}^{\star}\xi_n$$

Isolated (independent) models

$$\boldsymbol{X}_n^{\mathrm{A}} = \boldsymbol{f}\left(\boldsymbol{X}_{n-1}^{\mathrm{A}}, \ldots, \boldsymbol{X}_{n-L}^{\mathrm{A}}\right) + \widehat{\boldsymbol{g}}^{\star}\xi_n$$

$$\boldsymbol{X}_n^{\mathrm{I}} = \boldsymbol{f}\left(\boldsymbol{X}_{n-1}^{\mathrm{I}}, \ldots, \boldsymbol{X}_{n-L}^{\mathrm{I}}\right) + \widehat{\boldsymbol{g}}^{\star}\xi_n$$

$$\boldsymbol{X}_n^{\mathrm{P}} = \boldsymbol{f}\left(\boldsymbol{X}_{n-1}^{\mathrm{P}}, \ldots, \boldsymbol{X}_{n-L}^{\mathrm{P}}\right) + \widehat{\boldsymbol{g}}^{\star}\xi_n$$

The method of considering interaction between basins

Joint model

A **single** (joint) **model** for all basins.
Data from all basins are combined into one dataset.

Joint model

$$\boldsymbol{X}^{*} = (\boldsymbol{X}^{\mathrm{Atlantic}}, \boldsymbol{X}^{\mathrm{Indian}}, \boldsymbol{X}^{\mathrm{Pacific}}), \quad \boldsymbol{X}_n^{*} = \boldsymbol{f}(\boldsymbol{X}_{n-1}^{*}, \dots, \boldsymbol{X}_{n-L}^{*}) + \hat{\boldsymbol{g}}^{*}\boldsymbol{\xi}_n$$

Isolated (independent) models

$$\boldsymbol{X}_n^{\mathrm{A}} = \boldsymbol{f}\left(\boldsymbol{X}_{n-1}^{\mathrm{A}}, \dots, \boldsymbol{X}_{n-L}^{\mathrm{A}}\right) + \hat{\boldsymbol{g}}^{*}\boldsymbol{\xi}_n$$

$$\boldsymbol{X}_n^{\mathrm{I}} = \boldsymbol{f}\left(\boldsymbol{X}_{n-1}^{\mathrm{I}}, \dots, \boldsymbol{X}_{n-L}^{\mathrm{I}}\right) + \hat{\boldsymbol{g}}^{*}\boldsymbol{\xi}_n$$

$$\boldsymbol{X}_n^{\mathrm{P}} = \boldsymbol{f}\left(\boldsymbol{X}_{n-1}^{\mathrm{P}}, \dots, \boldsymbol{X}_{n-L}^{\mathrm{P}}\right) + \hat{\boldsymbol{g}}^{*}\boldsymbol{\xi}_n$$

Results

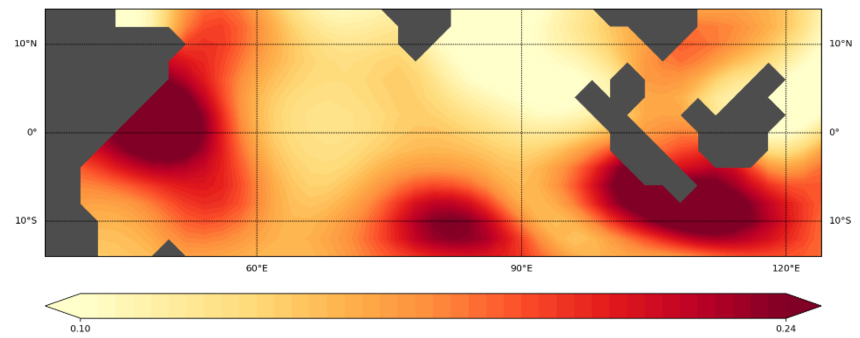
Data: ERSST version 5

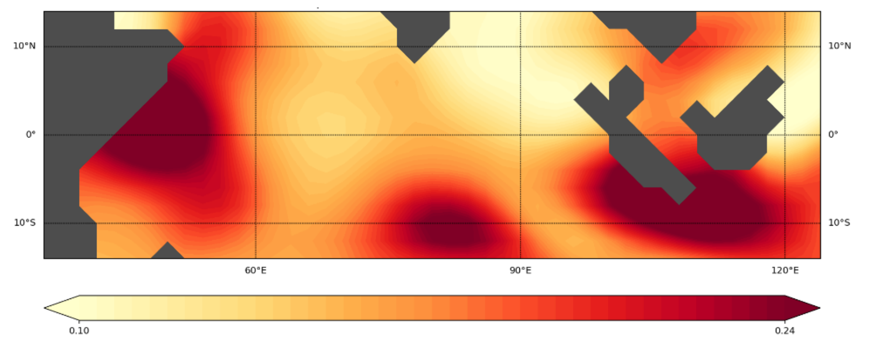
The variance of one-step forecast error

Joint model of three oceans

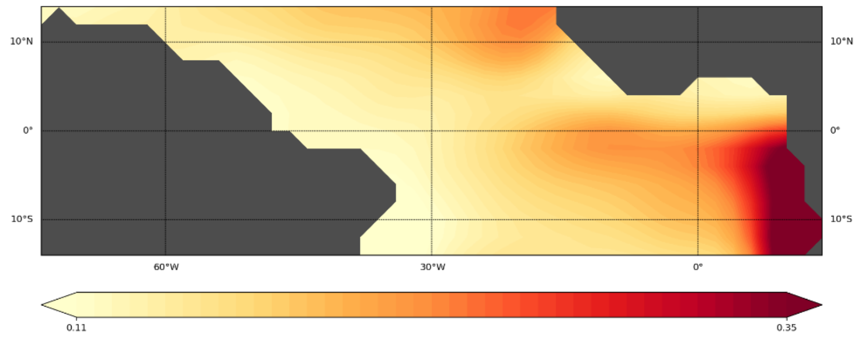
Isolated models for three oceans

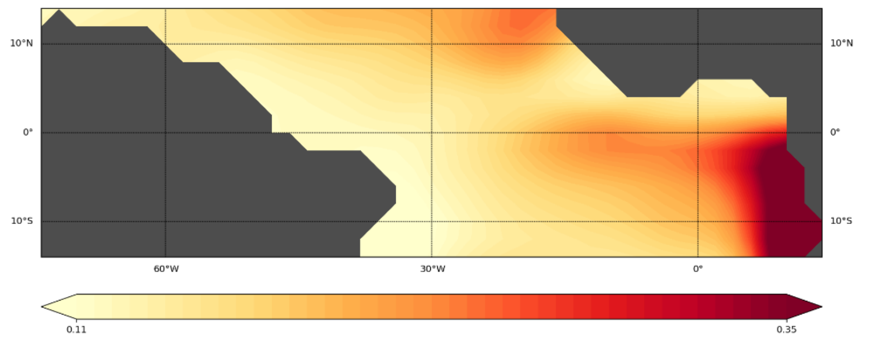
Indian ocean



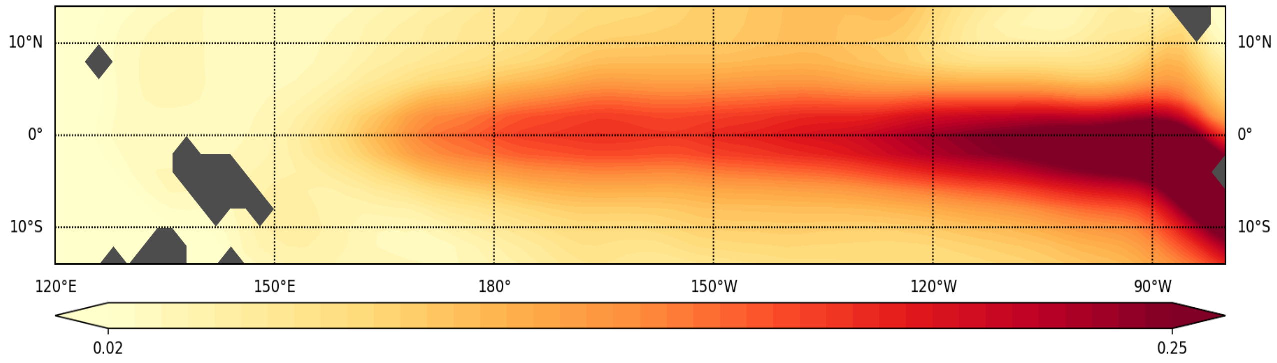


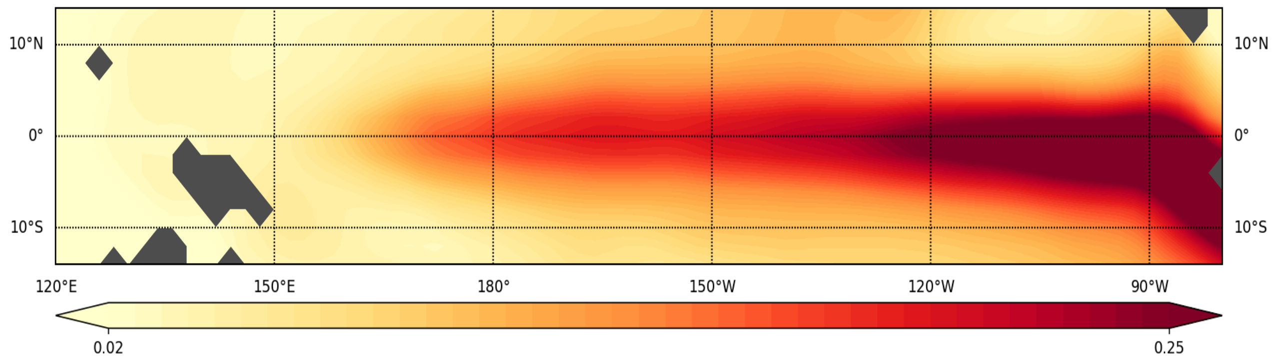
Atlantic ocean





Pacific ocean



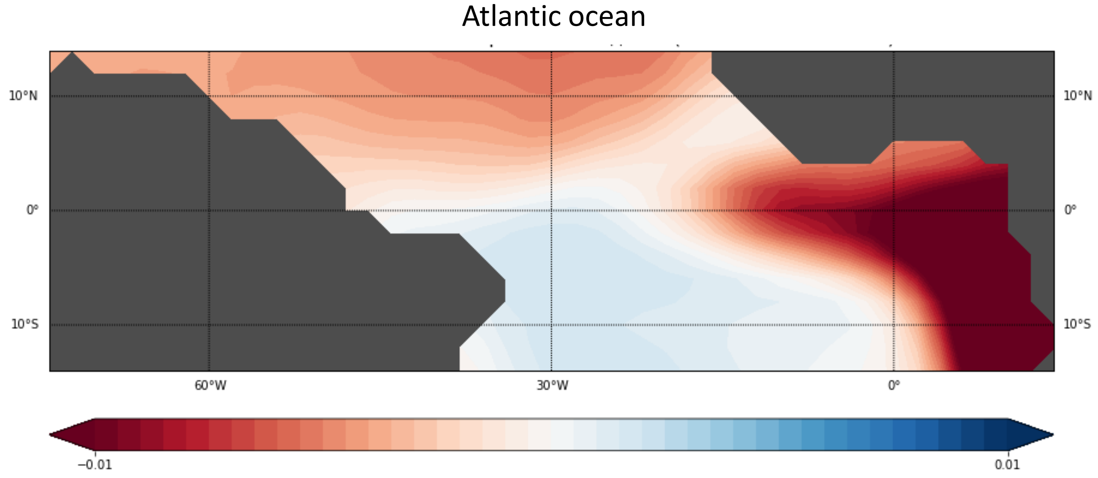


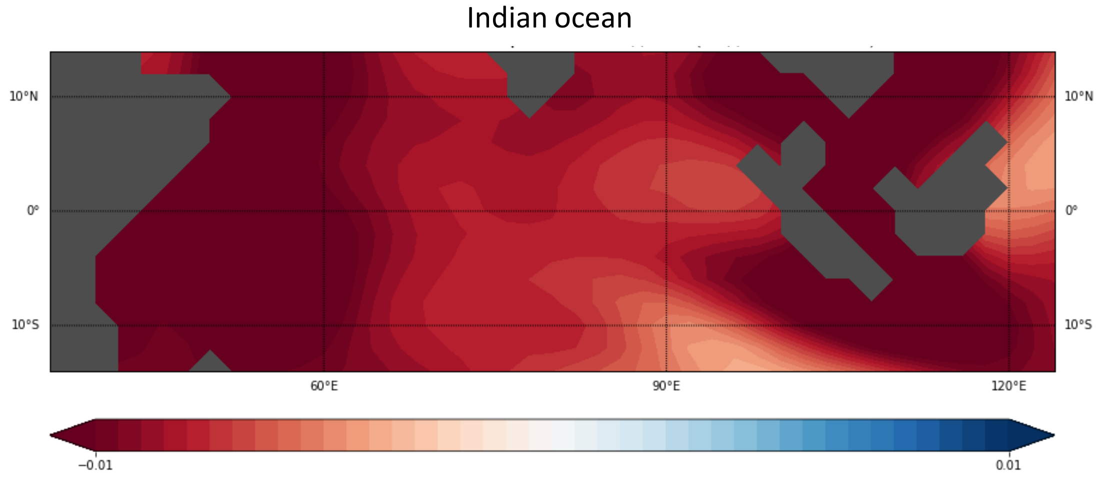
Results

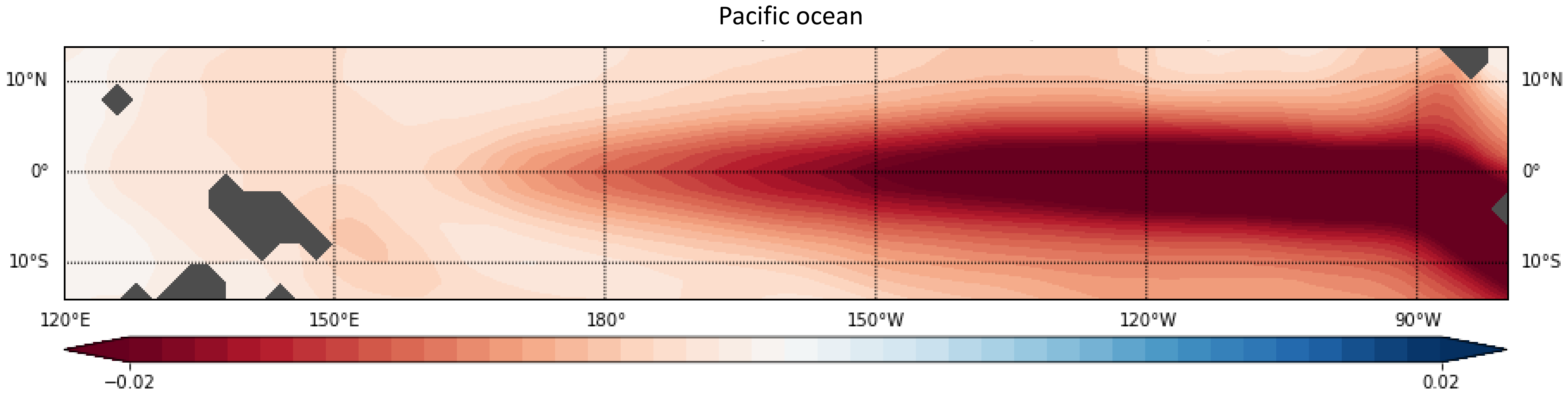
Improving the forecast

<u>Data: ERSST version 5</u>

The difference between forecast error variances of joint model and isolated models





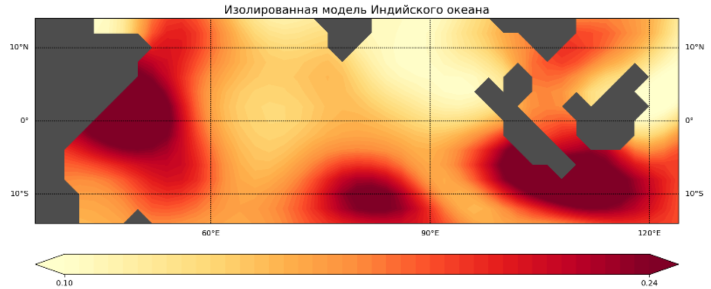


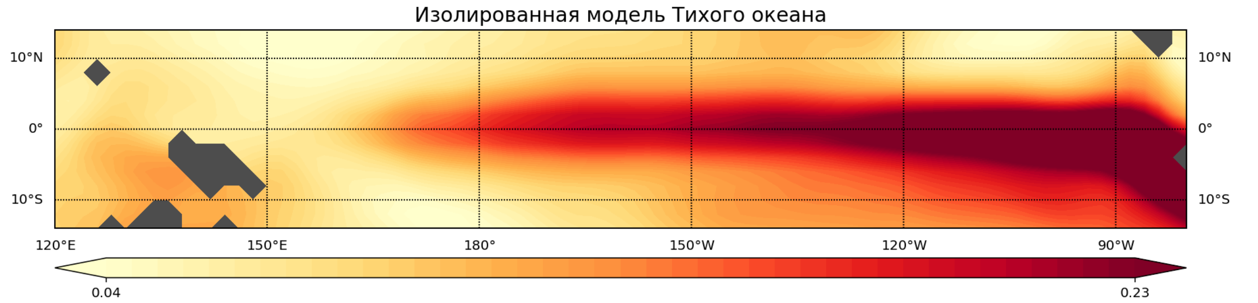
Results

Improving the forecast

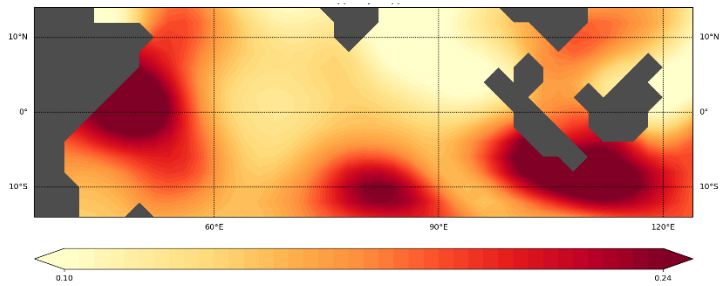
Data: ERSST version 5

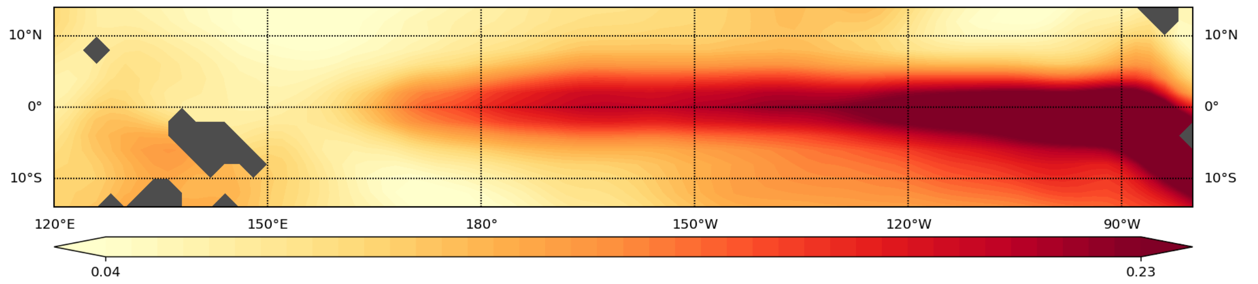
The variance of one-step forecast error





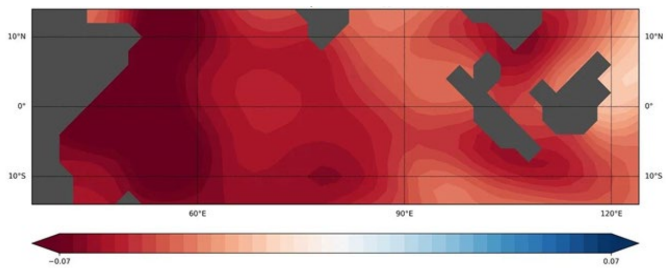
Isolated model

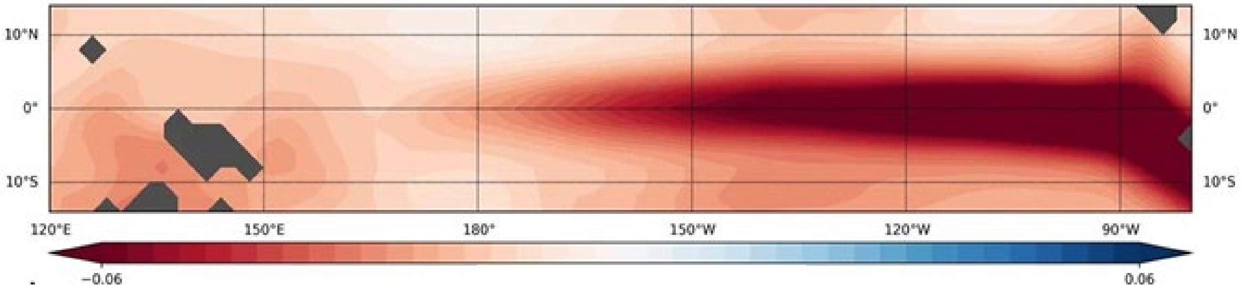




Joint model

The difference between forecast error variances of joint model and isolated models





Results

Comparison of methods for accounting for connectivity

Data: ERSST version 5

Comparison of forecast quality between joint model and "forced" models

Joint model

$$\mathbf{X}^*_n = f(\mathbf{X}^*_{n-1}, \dots, \mathbf{X}^*_{n-L}; t_n) + \hat{\mathbf{g}}\, \boldsymbol{\xi}_n$$

$\mathbf{X}^*$ Dataset from all basins

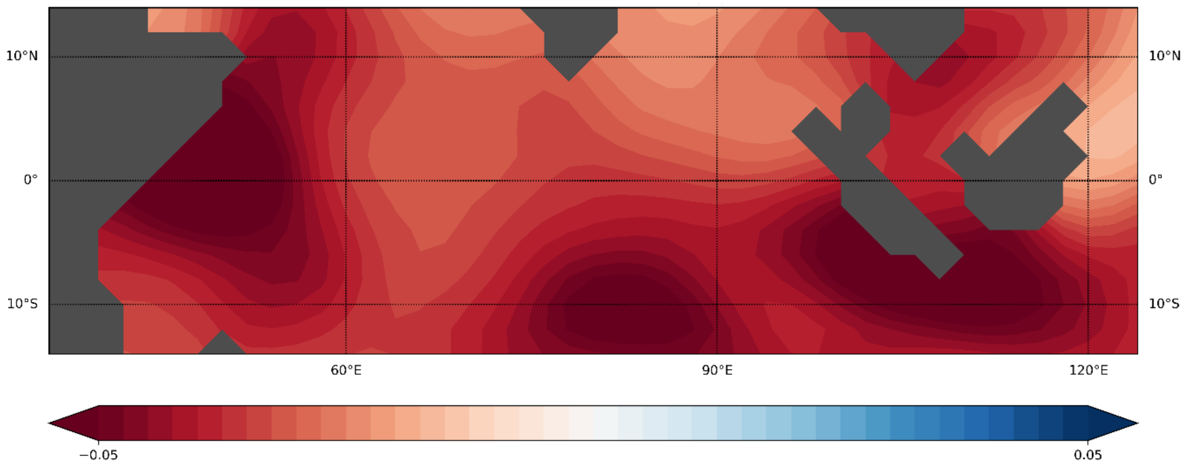
Forced models

$$\mathbf{X}_n = f(\mathbf{X}_{n-1}, \dots, \mathbf{X}_{n-L}; \mathbf{F}_n, t_n) + \hat{\mathbf{g}}\, \boldsymbol{\xi}_n$$

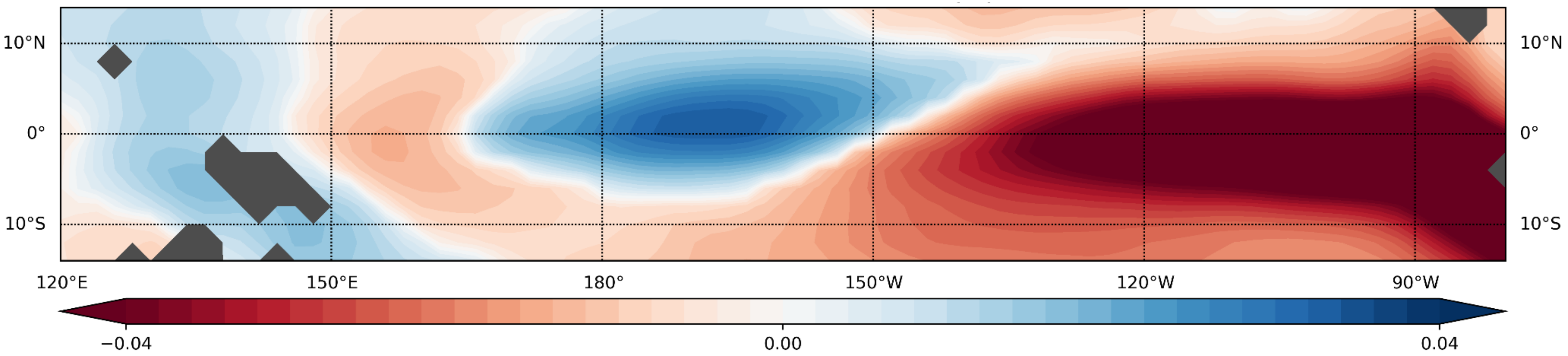
$\mathbf{X}$ Time series from basin of interest

$\mathbf{F}$ Time series from other basin

Indian ocean



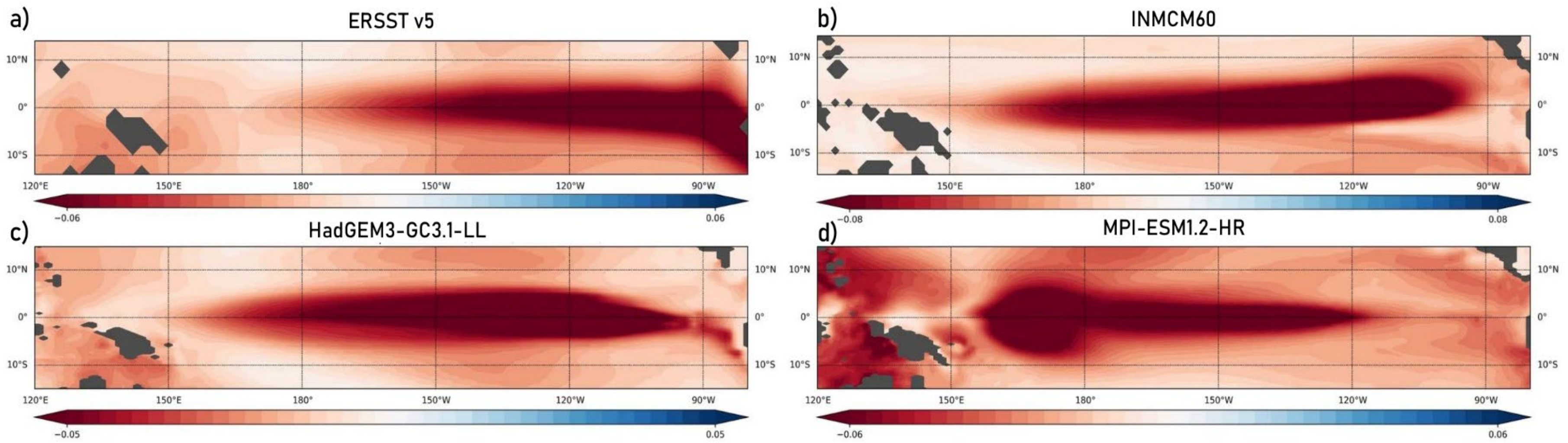
Pacific ocean



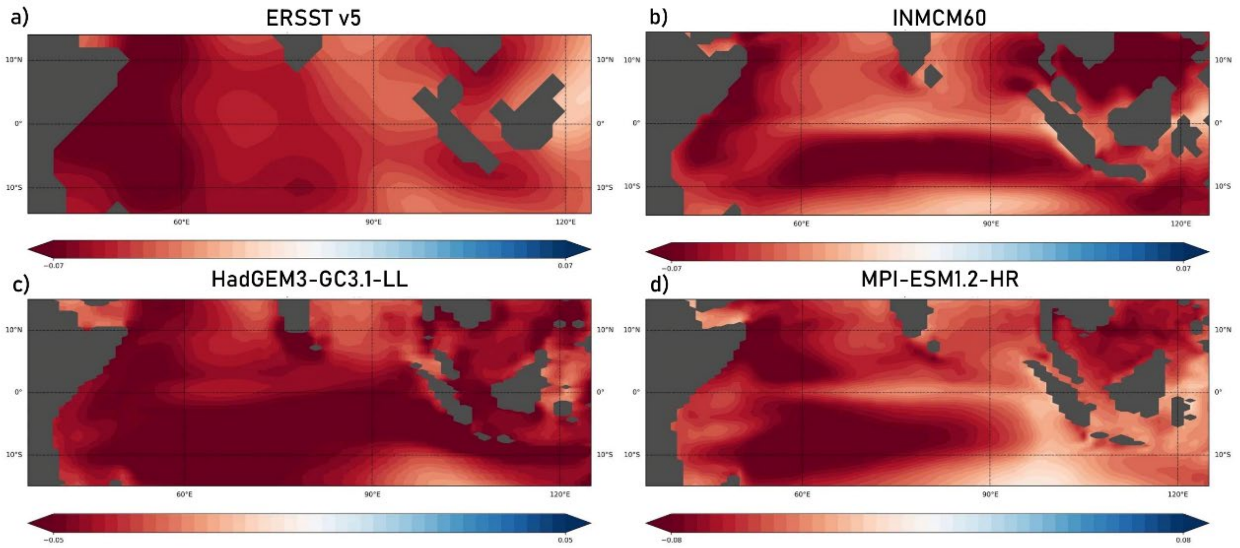
Results

Improved forecast due to accounting for connectivity

Pacific ocean



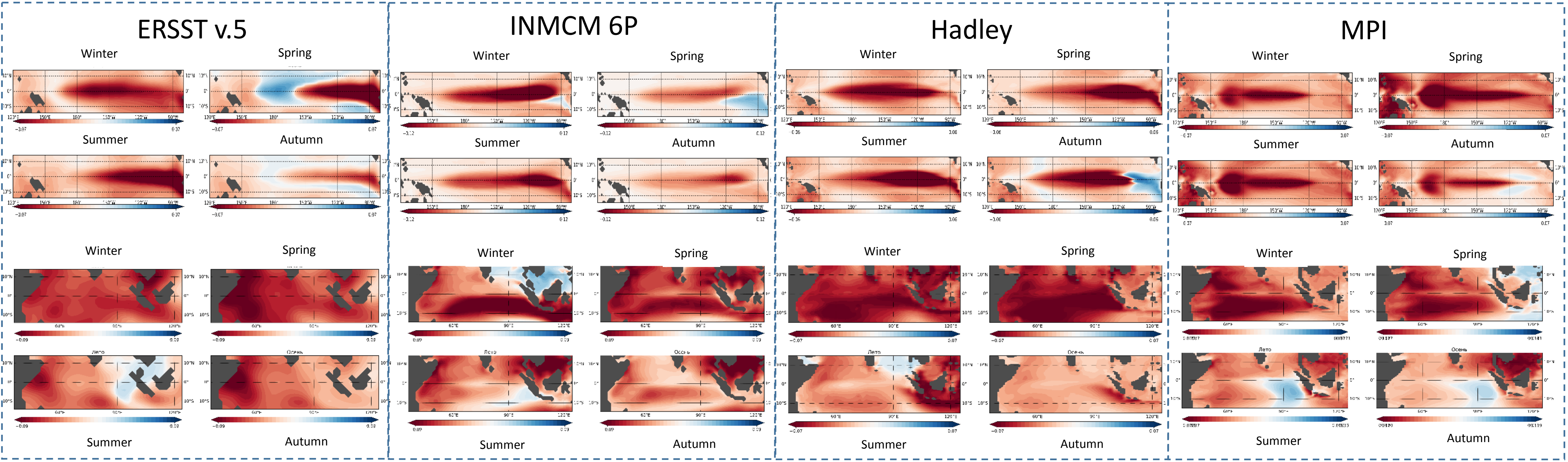
Indian ocean



Results

Improved forecast due to accounting for connectivity

Seasonal forecast improvement



Results

Improved forecast due to accounting for connectivity

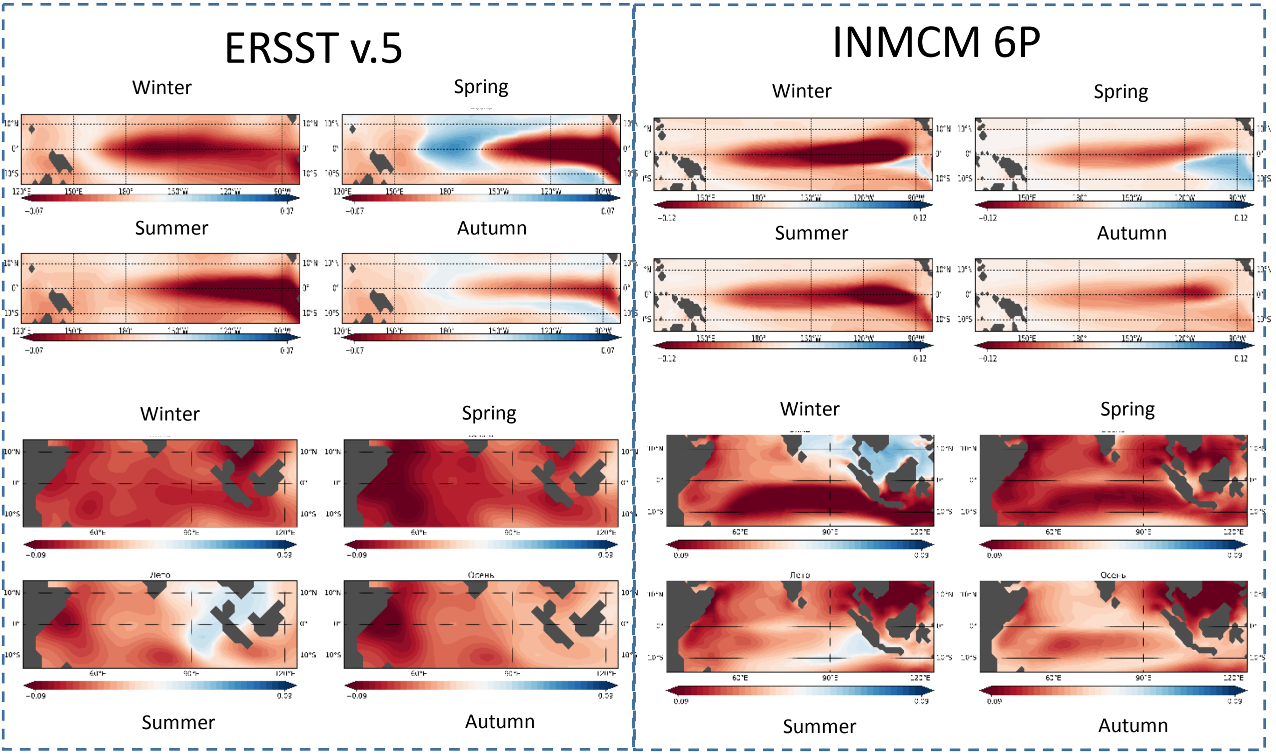
Pacific ocean

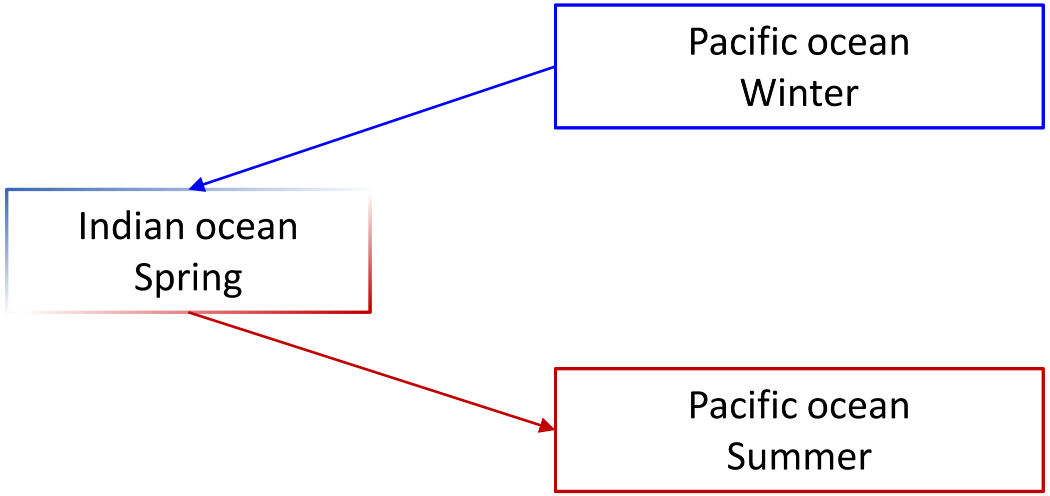
Data	Winter	Spring	Summer	Autumn
ERSST v5	-0.03	-0.02	-0.03	-0.01
INMCM 6P (2023)	-0.05	-0.02	-0.04	-0.03

Indian ocean

Data	Winter	Spring	Summer	Autumn
ERSST v5	-0.06	-0.07	-0.03	-0.05
INMCM 6P (2023)	-0.04	-0.06	-0.05	-0.04

Seasonal forecast improvement





Conclusions

- Taking into account the relationship between the basins of the Indian and Pacific Oceans has led to a significant improvement in the forecast of SST anomalies.

- Of the two considered methods of considering interaction (impacts) between basins, the method of constructing joint models works better.

- In all the considered models and reanalysis data, significantly different patterns of forecast improvement are found, allowing us to speak only about qualitative similarity in some cases.

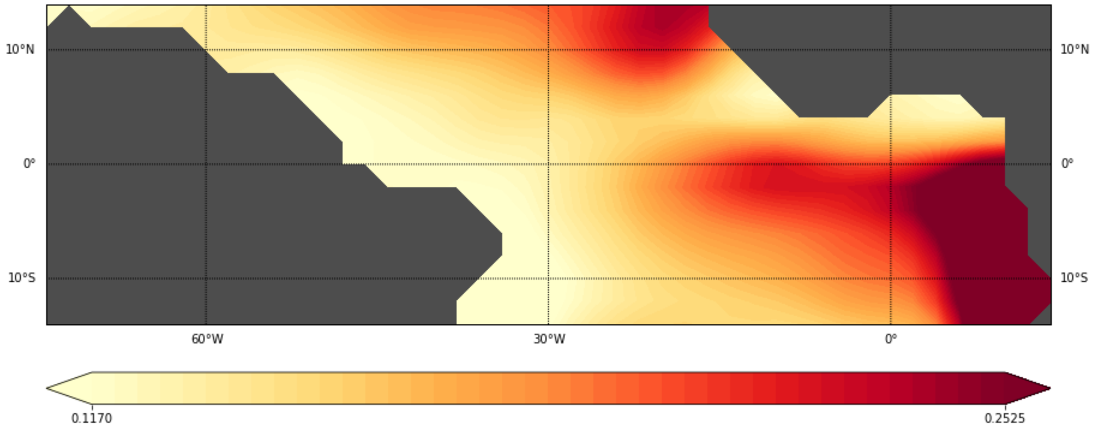
Results

<u>Data: ERSST version 5</u>

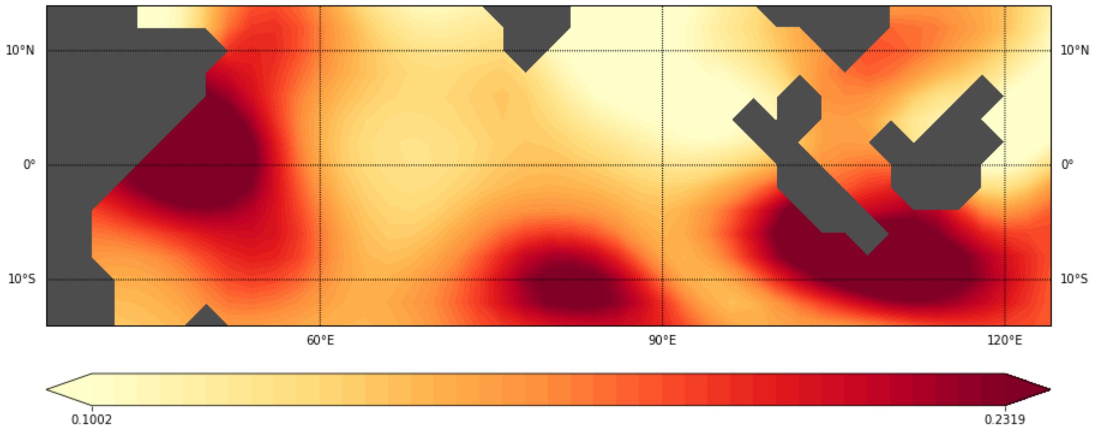
Joint model of three oceans

The variance of one-step forecast error

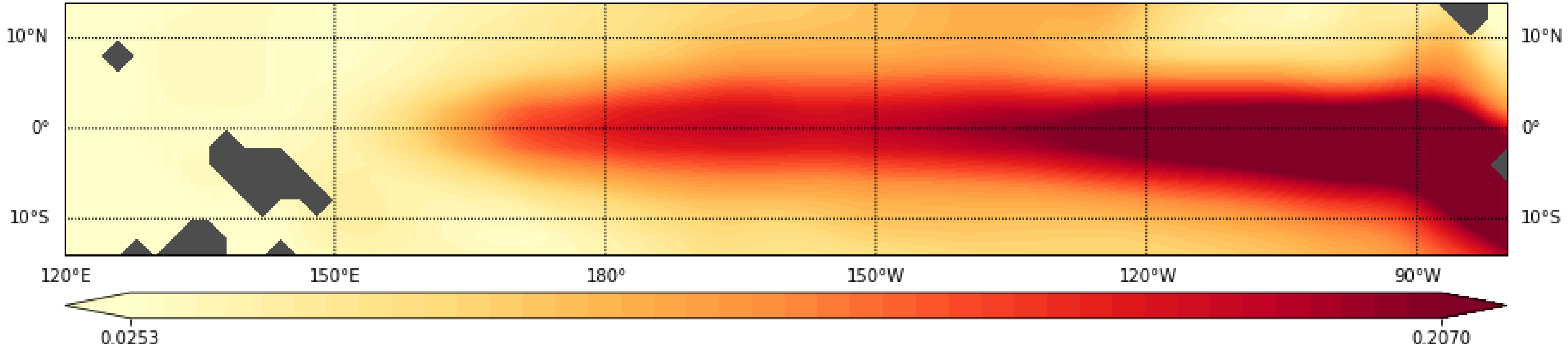
Joint model , Atlantic ocean



Joint model , Indian ocean



Joint model , Pacific ocean



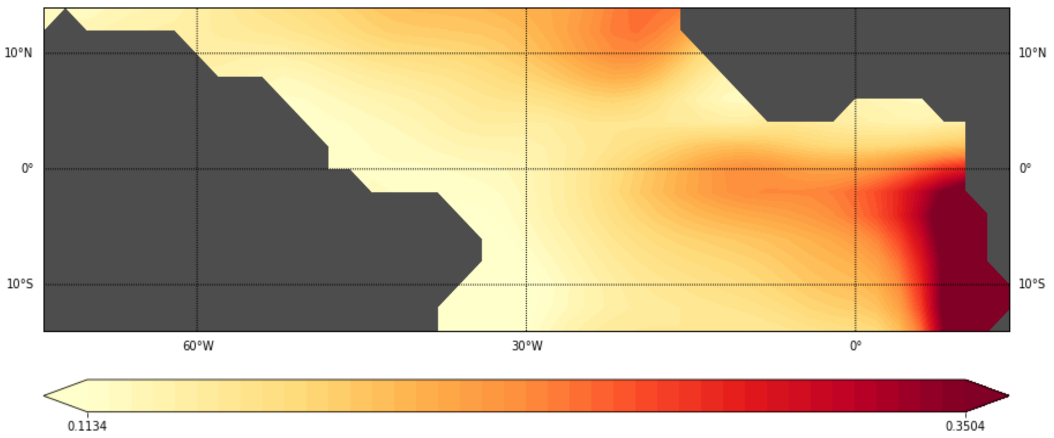
Results

Data: ERSST version 5

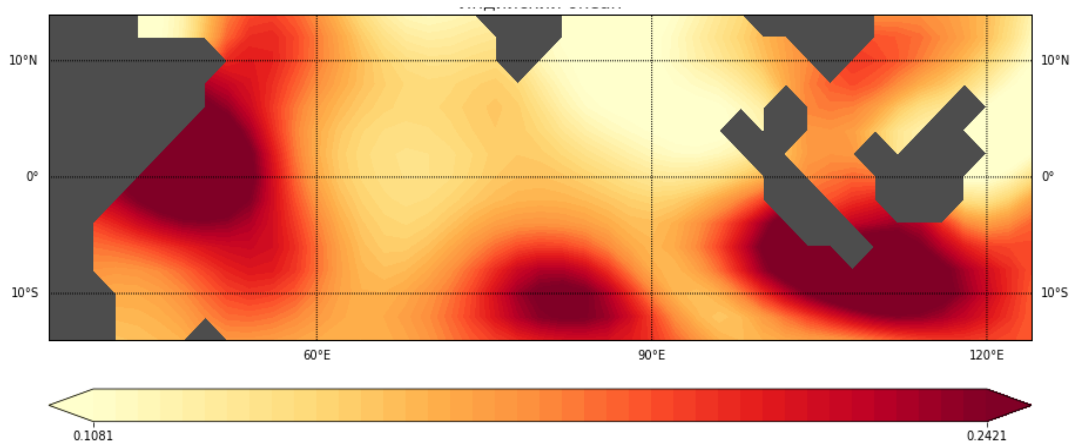
Isolated models for three oceans

The variance of one-step forecast error

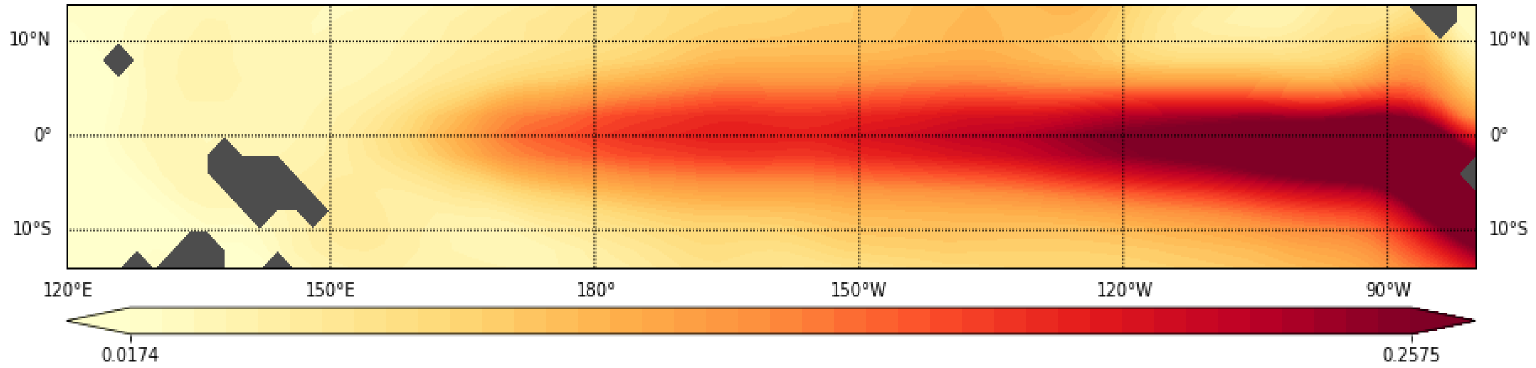
Isolated model of Atlantic ocean



Isolated model of Indian ocean



Isolated model of Pacific ocean



Empirical analysis: construction of data-driven (empirical) models

Model learning and optimization

Model $\boldsymbol{H_i}$

parameters $\boldsymbol{\mu} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\nu}, \boldsymbol{\gamma}, \hat{\mathbf{g}})$

hyperparameters $(L, \mathrm{m}_f)_i$

Bayes' theorem

$$P(H_i|\overline{X}) = \frac{P(\overline{X}|H_i) * P(H_i)}{\sum_i P(\overline{X}|H_i) * P(H_i)}$$

$$P(\overline{X}|H_i) = \int P(\overline{X}, \boldsymbol{\mu}|H_i) d\boldsymbol{\mu} = \int P(\overline{X}|\boldsymbol{\mu}, H_i) P(\boldsymbol{\mu}|H_i) d\boldsymbol{\mu} \xrightarrow{H_i} \max$$

Likelihood Prior

Laplace's method

L=-ln $P(\overline{X}|H_i)$ Bayesian evidence of the model H_i

$P(\overline{X}|\boldsymbol{\mu}, H_i) P(\boldsymbol{\mu}|H_i) = \exp(-\Psi_{\mathbf{H}_i}(\boldsymbol{\mu}))$

$\boldsymbol{\mu}^* = \underset{\boldsymbol{\mu}}{\operatorname{argmax}} P(\mathbf{x}|\boldsymbol{\mu}, H_i)$

$$L = -\ln P(\overline{X}|H_i) = \Psi_{\mathbf{H}_i}(\boldsymbol{\mu}^*) + \frac{1}{2} \ln \left[\left| \frac{1}{2\pi} \nabla \nabla^T \Psi_{\mathbf{H}_i}(\boldsymbol{\mu}^*) \right| \right]$$

"quality term" (quality of the data description by the model)

"penalty term" (penalty for the complexity of the model)