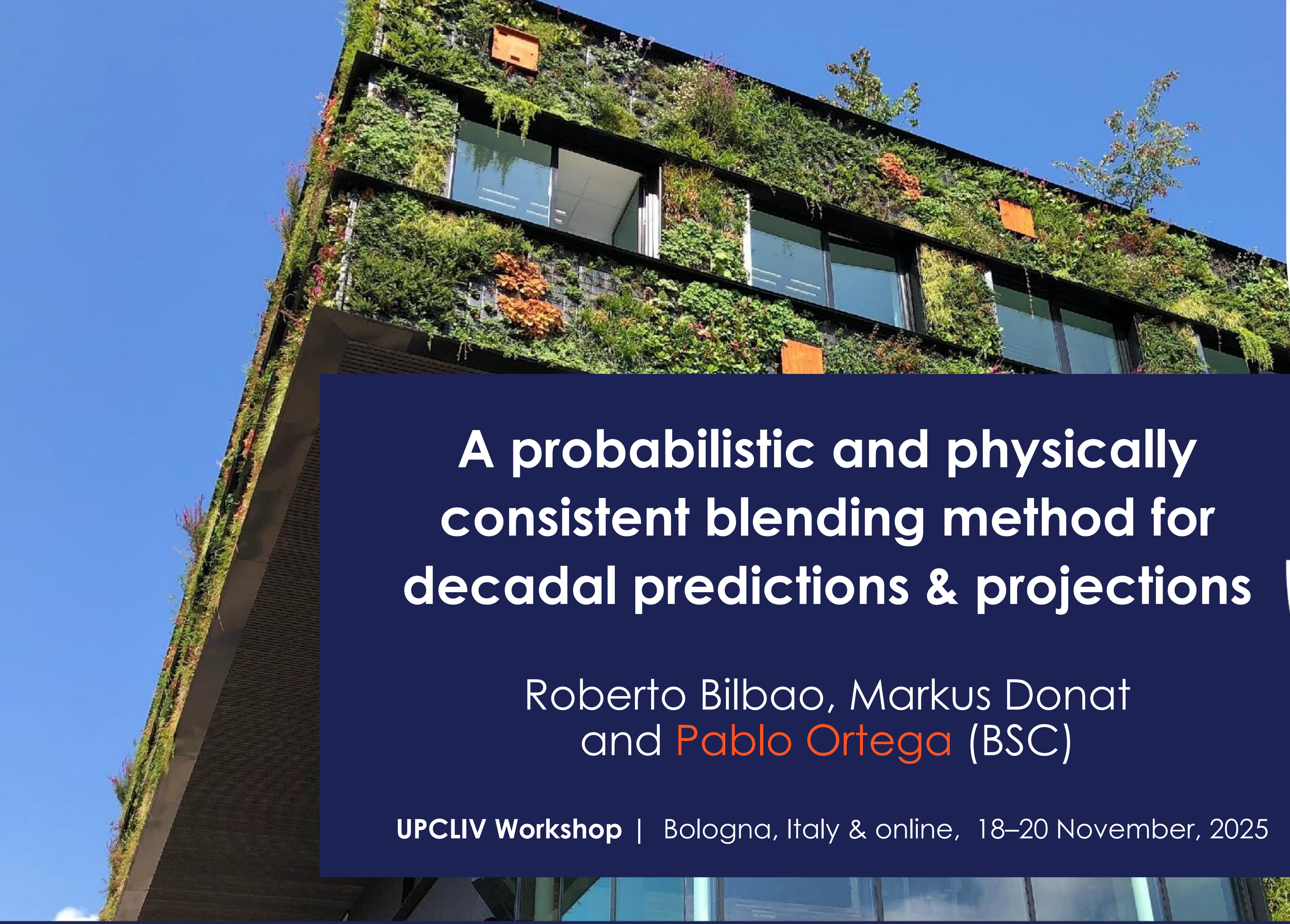


IMPETUS 4CHANGE

# A probabilistic and physically consistent blending method for decadal predictions & projections

Roberto Bilbao, Markus Donat
and Pablo Ortega (BSC)

**UPCLIV Workshop** | Bologna, Italy & online, 18–20 November, 2025

impetus4change.eu @I4C_eu

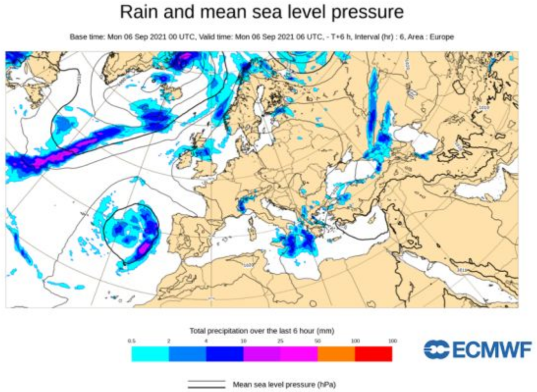
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081555

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# MOTIVATION

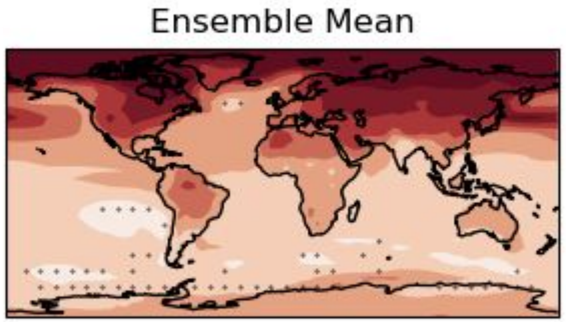
Climate Services

Stakeholders




ECMWF

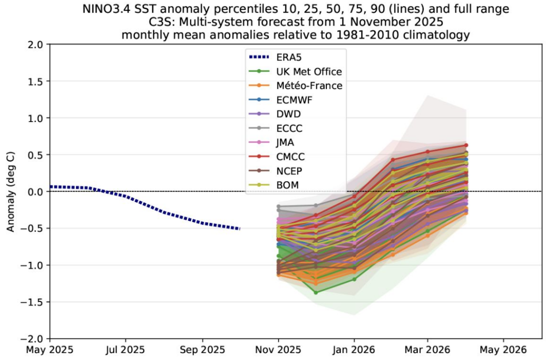
2024 predictions for 2025-2029 near-surface temperature



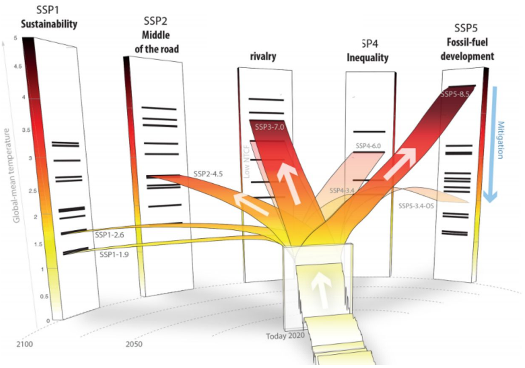

WMO Lead Centre for
Annual-to-Decadal Climate Prediction

| Weather forecasts | Subseasonal Predictions | Seasonal Predictions | Decadal Climate Predictions | Climate Projections |
|---|---|---|---|---|

Days — Weeks — Months — Years — Decades — Centuries



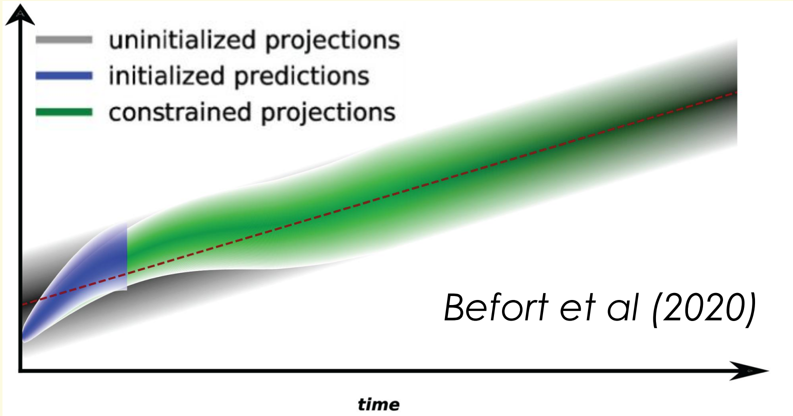

Copernicus Climate Change Service — Implemented by ECMWF




*Meinshausen et al. (2020)*

# SEAMLESS PREDICTION METHODS

## Constraining methods



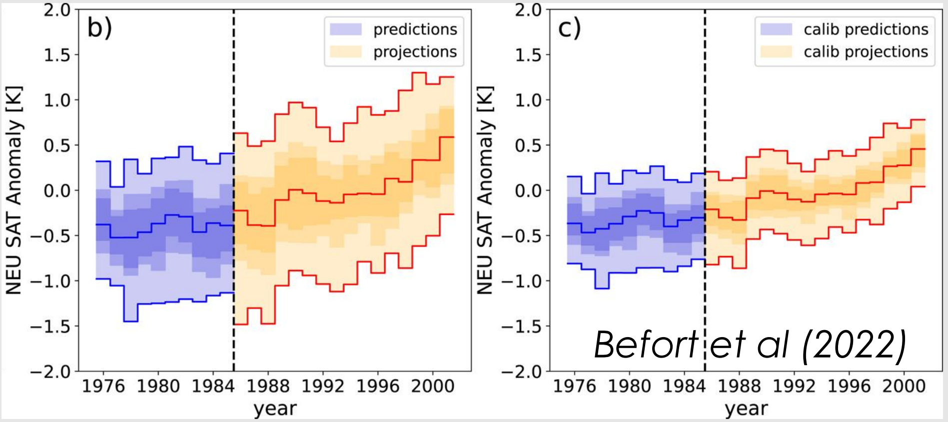

### Advantages

- Can boost predictive skill in the projections
- Constraints based on 1 variable but affect all system

### Limitations

- Currently work only with ensemble prediction means
- Probabilistic consistency is currently neglected

## Blending (via stitching)




### Advantages

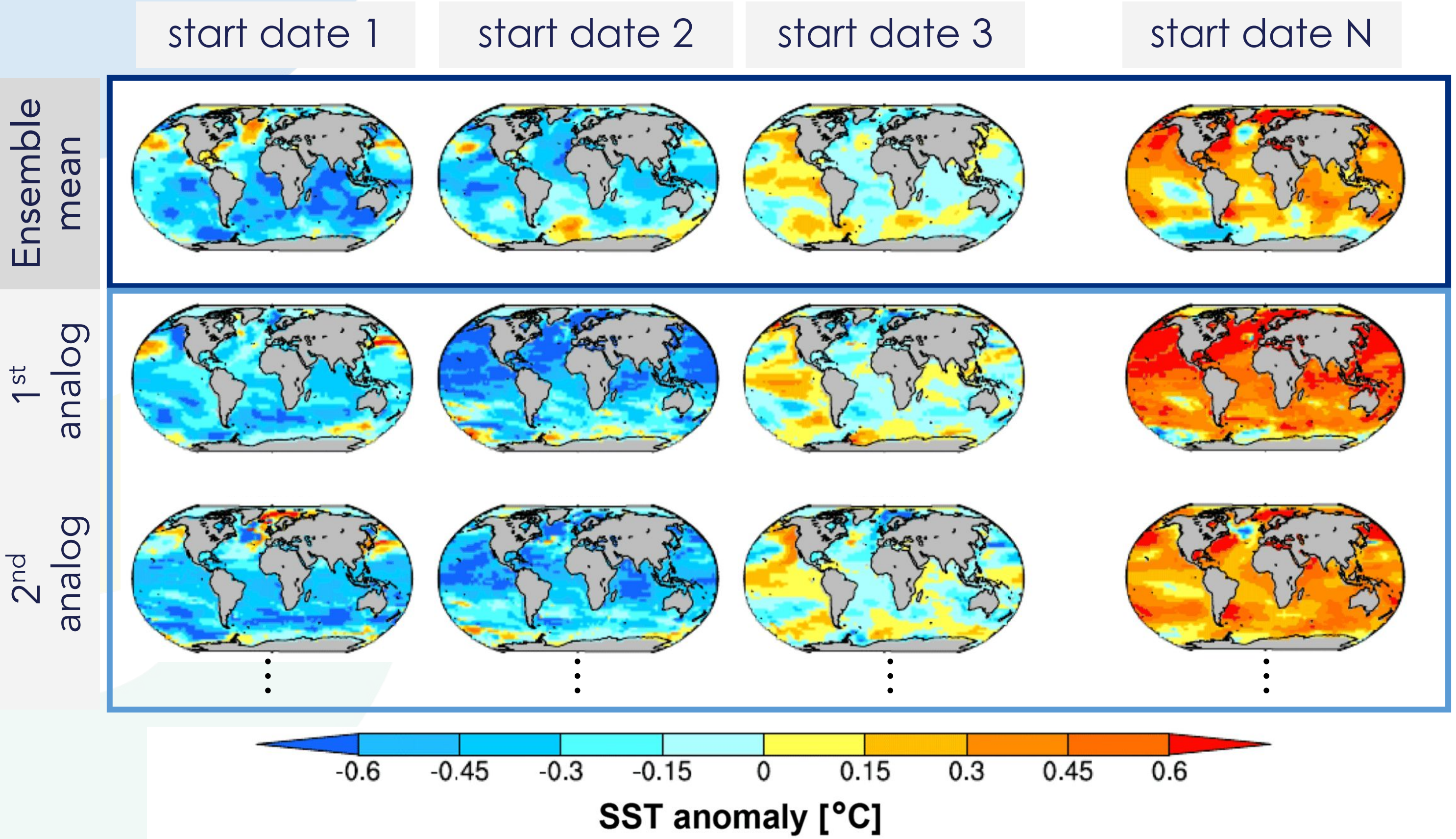
- Exploit full ensemble information
- Bias and variance correction largely improve reliability

### Limitations

- Do not address physical consistency
- Cannot be applied uniformly across variables
- Variance inflation not usable in non-gaussian variables

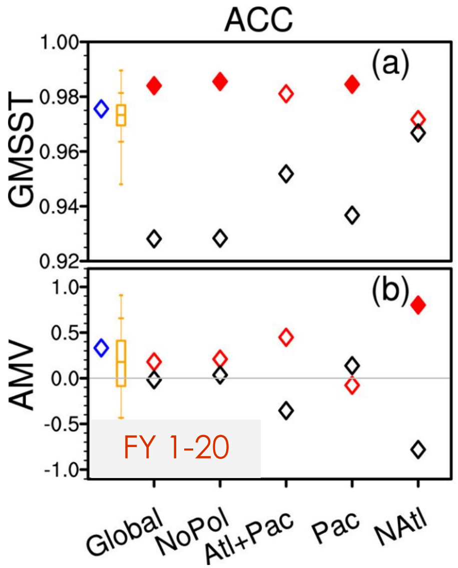
# SPATIAL ANALOGS – A METRIC FOR CONSTRAINING

A methodology used to constrain future projections with decadal predictions or observations




Decadal **Predictions**

Ensemble of **Projections**




This method has been applied to **different domains** to identify analogs for the **ensemble mean**

It helps retaining **predictability from internal variability sources**

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

*Mahmood et al. (2021, 2022)*

# SPATIAL ANALOGS – A METRIC ALSO FOR BLENDING

## Decadal Prediction Systems (DPS)

**133** members of **12** different DPS

| Model | Members |
|---|---|
| ~~BCC-CSM2-MR~~ | ~~8~~ |
| CanESM5-p2 | 20 |
| CESM1-1-CAM5-CMIP5 | 20 |
| CMCC-ESM2 | 10 |
| EC-Earth3 i1 | 10 |
| EC-Earth3 i2 | 10 |
| FGOALS-f3-L | 3 |
| HadGEM3-GC3.1-MM | 10 |
| IPSL-CM6A-LR | 10 |
| MIROC6 | 10 |
| MPI-ESM1-2-HR | 10 |
| NorCMP1 i1 | 10 |
| NorCMP1 i2 | 10 |

Blending after forecast **year 10**

## Historical Simulations

**364** historical simulations from **29** different models

| Model | Members | Model | Members |
|---|---|---|---|
| ACCESS-CM2 | 5 | FIO-ESM-2-0 | 3 |
| BCC-CSM2-MR | 3 | GISS-E2-1-G | 8 |
| CAMS-CSM1-0 | 2 | HadGEM3-GC31-LL | 4 |
| CanESM5-p1 | 25 | IPSL-CM6A-LR | 33 |
| CanESM5-p2 | 40 | KACE-1-0-G | 3 |
| CAS-ESM2-0 | 2 | MIROC-ES2L | 30 |
| CESM2-WACCM | 3 | MIROC6 | 50 |
| CMCC-CM2-SR5 | 10 | MPI-ESM1-2-HR | 10 |
| CMCC-ESM2 | 1 | MPI-ESM1-2-LR | 10 |
| CNRM-CM6-1 | 30 | MRI-ESM2-0 | 10 |
| CNRM-ESM2-1 | 11 | NESM3 | 5 |
| EC-Earth3 | 22 | NorESM2-LM | 3 |
| FGOALS-f3-L | 3 | NorESM2-MM | 1 |
| FGOALS-g3 | 4 | NorCMP1 | 30 |

# SPATIAL ANALOGS – A METRIC ALSO FOR BLENDING

## Decadal Prediction Systems (DPS)

**133** members of **12** different DPS

| Model | Members |
|---|---|
| ~~BCC-CSM2-MR~~ | ~~8~~ |
| CanESM5-p2 | 20 |
| CESM1-1-CAM5-CMIP5 | 20 |
| CMCC-ESM2 | 10 |
| EC-Earth3 i1 | 10 |
| EC-Earth3 i2 | 10 |
| FGOALS-f3-L | 3 |
| HadGEM3-GC3.1-MM | 10 |
| IPSL-CM6A-LR | 10 |
| MIROC6 | 10 |
| MPI-ESM1-2-HR | 10 |
| NorCMP1 i1 | 10 |
| NorCMP1 i2 | 10 |

Blending after forecast **year 10**

Analogs are identified for each prediction member seeking to maintain **physical and statistical coherence**

## Historical Simulations

**364** historical simulations from **29** different models

| Model | Members | Model | Members |
|---|---|---|---|
| ACCESS-CM2 | 5 | FIO-ESM-2-0 | 3 |
| BCC-CSM2-MR | 3 | GISS-E2-1-G | 8 |
| CAMS-CSM1-0 | 2 | HadGEM3-GC31-LL | 4 |
| CanESM5-p1 | 25 | IPSL-CM6A-LR | 33 |
| CanESM5-p2 | 40 | KACE-1-0-G | 3 |
| CAS-ESM2-0 | 2 | MIROC-ES2L | 30 |
| CESM2-WACCM | 3 | MIROC6 | 50 |
| CMCC-CM2-SR5 | 10 | MPI-ESM1-2-HR | 10 |
| CMCC-ESM2 | 1 | MPI-ESM1-2-LR | 10 |
| CNRM-CM6-1 | 30 | MRI-ESM2-0 | 10 |
| CNRM-ESM2-1 | 11 | NESM3 | 5 |
| EC-Earth3 | 22 | NorESM2-LM | 3 |
| FGOALS-f3-L | 3 | NorESM2-MM | 1 |
| FGOALS-g3 | 4 | NorCMP1 | 30 |

# SPATIAL ANALOGS – A METRIC ALSO FOR BLENDING

## Decadal Prediction Systems (DPS)

**133** members of **12** different DPS

| Model | Members |
|---|---|
| ~~BCC-CSM2-MR~~ | ~~8~~ |
| CanESM5-p2 | 20 |
| CESM1-1-CAM5-CMIP5 | 20 |
| CMCC-ESM2 | 10 |
| EC-Earth3 i1 | 10 |
| EC-Earth3 i2 | 10 |
| FGOALS-f3-L | 3 |
| HadGEM3-GC3.1-MM | 10 |
| IPSL-CM6A-LR | 10 |
| MIROC6 | 10 |
| MPI-ESM1-2-HR | 10 |
| NorCMP1 i1 | 10 |
| NorCMP1 i2 | 10 |

Different methodological choices have been tested

| |
|---|
| **Metric to assess similarity:**<br>RMSE, Pattern Correlation and Taylor Skill Score |
| **Spatial Anomaly definition:**<br>Centered vs Uncentered |
| **Spatial domain:**<br>Global vs North Atlantic |
| **Ensemble member selection:**<br>with vs without repetition |
| **Historical and DPS model overlap:**<br>Full vs Partial |

## Historical Simulations

**364** historical simulations from **29** different models

| Model | Members | Model | Members |
|---|---|---|---|
| ACCESS-CM2 | 5 | FIO-ESM-2-0 | 3 |
| BCC-CSM2-MR | 3 | GISS-E2-1-G | 8 |
| CAMS-CSM1-0 | 2 | HadGEM3-GC31-LL | 4 |
| CanESM5-p1 | 25 | IPSL-CM6A-LR | 33 |
| CanESM5-p2 | 40 | KACE-1-0-G | 3 |
| CAS-ESM2-0 | 2 | MIROC-ES2L | 30 |
| CESM2-WACCM | 3 | MIROC6 | 50 |
| CMCC-CM2-SR5 | 10 | MPI-ESM1-2-HR | 10 |
| CMCC-ESM2 | 1 | MPI-ESM1-2-LR | 10 |
| CNRM-CM6-1 | 30 | MRI-ESM2-0 | 10 |
| CNRM-ESM2-1 | 11 | NESM3 | 5 |
| EC-Earth3 | 22 | NorESM2-LM | 3 |
| FGOALS-f3-L | 3 | NorESM2-MM | 1 |
| FGOALS-g3 | 4 | NorCMP1 | 30 |

# SPATIAL ANALOGS – A METRIC ALSO FOR BLENDING

## Decadal Prediction Systems (DPS)

**133** members of **12** different DPS

| Model | Members |
|---|---|
| ~~BCC-CSM2-MR~~ | ~~8~~ |
| CanESM5-p2 | 20 |
| CESM1-1-CAM5-CMIP5 | 20 |
| CMCC-ESM2 | 10 |
| EC-Earth3 i1 | 10 |
| EC-Earth3 i2 | 10 |
| FGOALS-f3-L | 3 |
| HadGEM3-GC3.1-MM | 10 |
| IPSL-CM6A-LR | 10 |
| MIROC6 | 10 |
| MPI-ESM1-2-HR | 10 |
| NorCMP1 i1 | 10 |
| NorCMP1 i2 | 10 |

Different methodological choices have been tested

**Metric to assess similarity:**
RMSE, Pattern Correlation and Taylor Skill Score

**Spatial Anomaly definition:**
Centered vs Uncentered

**Spatial domain:**
Global vs North Atlantic

**Ensemble member selection:**
with vs without repetition

**Historical and DPS model overlap:**
Full vs Partial

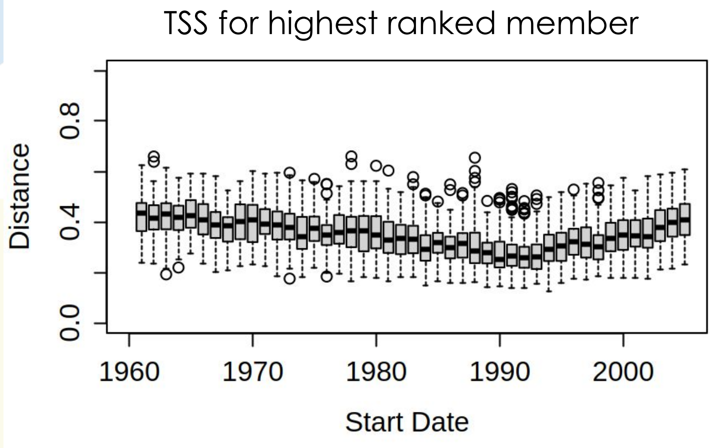
## Historical Simulations

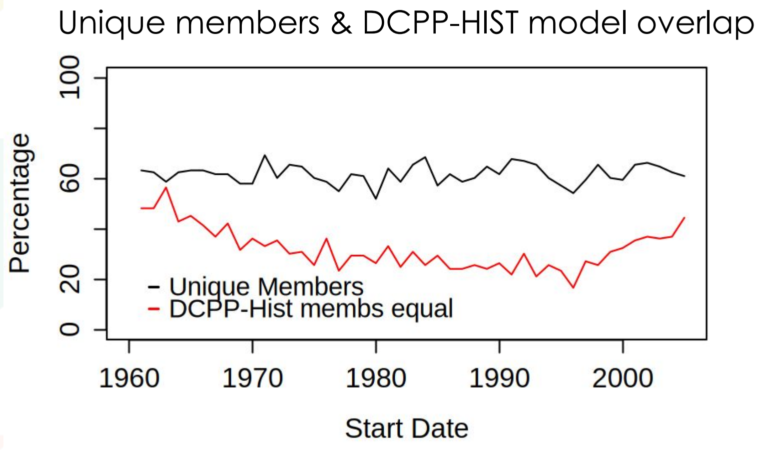
**364** historical simulations from **29** different models

| Model | Members | Model | Members |
|---|---|---|---|
| ACCESS-CM2 | 5 | FIO-ESM-2-0 | 3 |
| BCC-CSM2-MR | 3 | GISS-E2-1-G | 8 |
| CAMS-CSM1-0 | 2 | HadGEM3-GC31-LL | 4 |
| CanESM5-p1 | 25 | IPSL-CM6A-LR | 33 |
| CanESM5-p2 | 40 | KACE-1-0-G | 3 |
| CAS-ESM2-0 | 2 | MIROC-ES2L | 30 |
| CESM2-WACCM | 3 | MIROC6 | 50 |
| CMCC-CM2-SR5 | 10 | MPI-ESM1-2-HR | 10 |
| CMCC-ESM2 | 1 | MPI-ESM1-2-LR | 10 |
| CNRM-CM6-1 | 30 | MRI-ESM2-0 | 10 |
| CNRM-ESM2-1 | 11 | NESM3 | 5 |
| EC-Earth3 | 22 | NorESM2-LM | 3 |
| FGOALS-f3-L | 3 | NorESM2-MM | 1 |
| FGOALS-g3 | 4 | NorCMP1 | 30 |

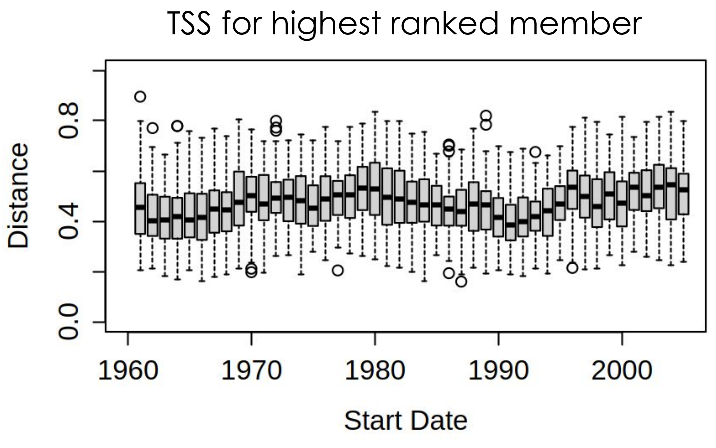
# TEMPORAL CONSISTENCY OF ANALOGS PER DPS

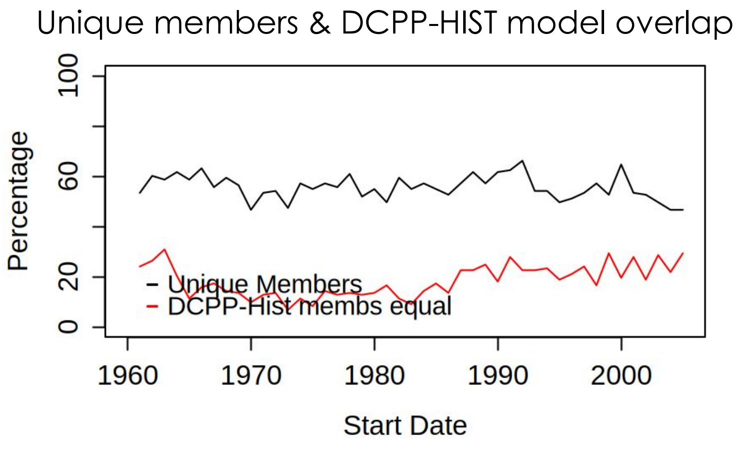
## Global domain (Centered Taylor Skill Score)



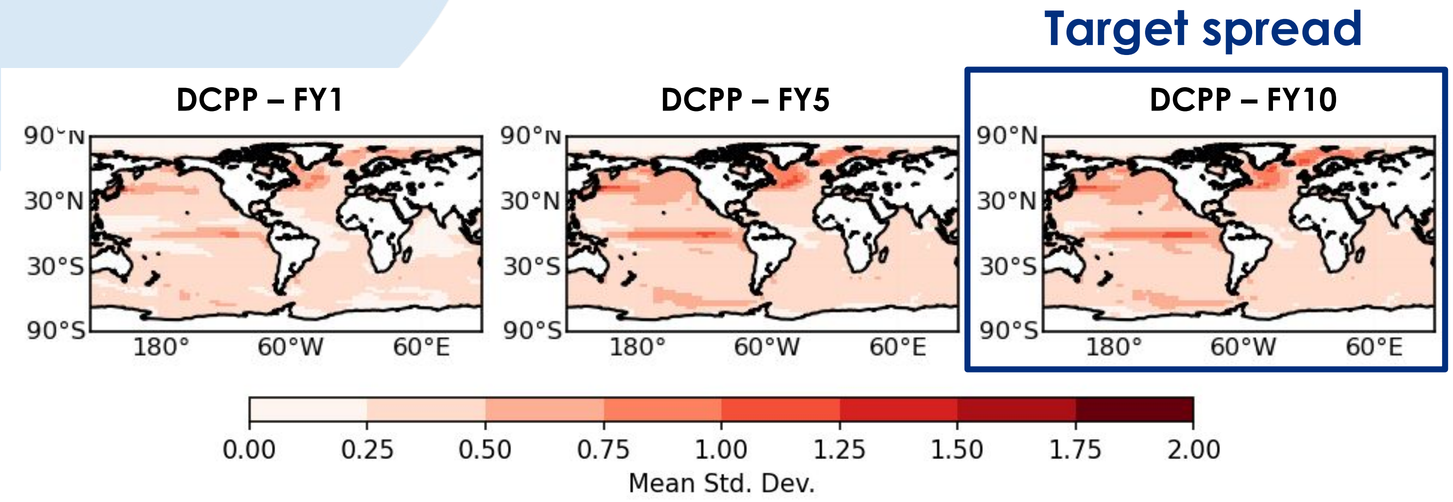

## North Atlantic domain (Centered Taylor Skill Score)



# IMPACT OF THE SELECTED APPROACH ON SPREAD

**Mean intra-ensemble spread** (STD in member dimension averaged across start dates)

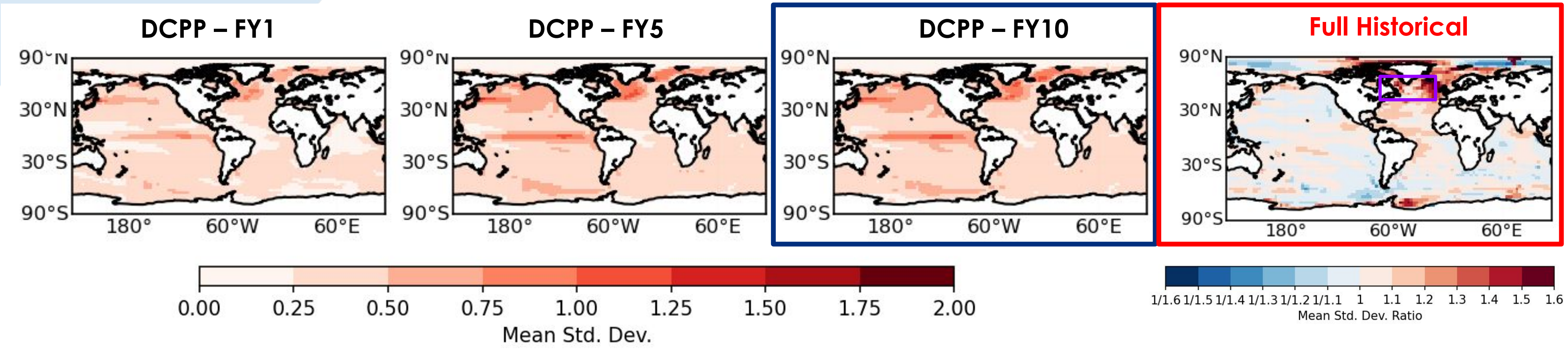


# IMPACT OF THE SELECTED APPROACH ON SPREAD

**Mean intra-ensemble spread** (STD in member dimension averaged across start dates)

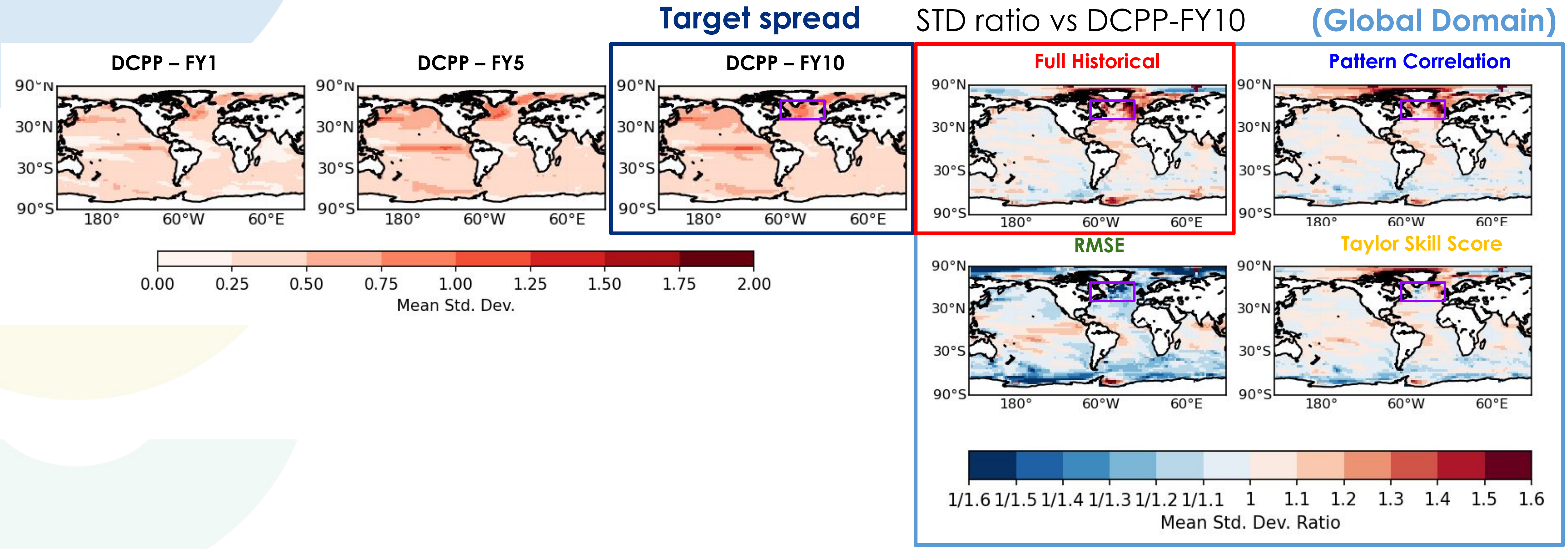
Target spread

STD ratio vs DCPP-FY10



# IMPACT OF THE SELECTED APPROACH ON SPREAD

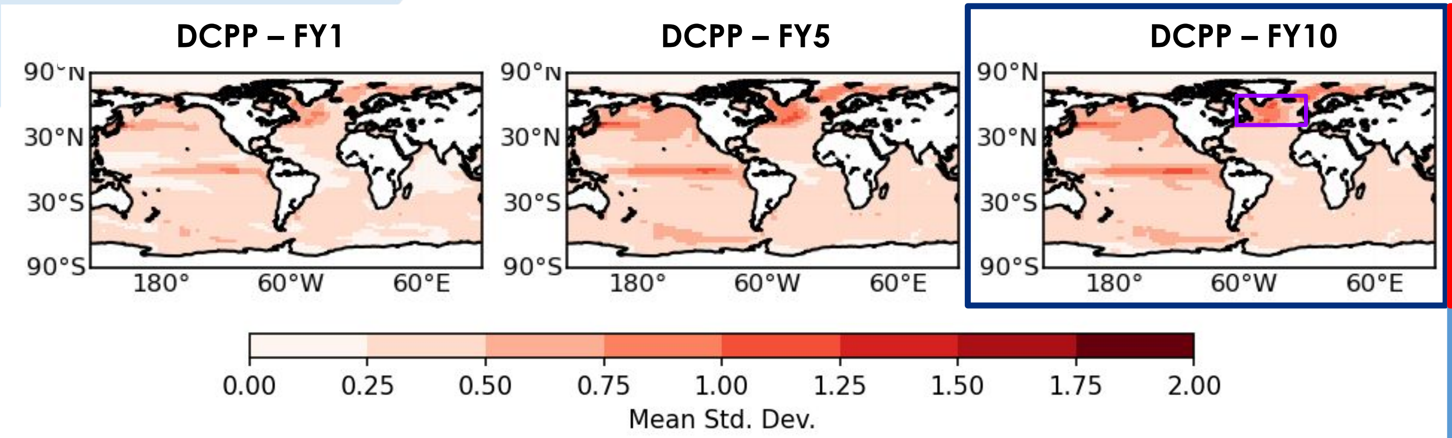
**Mean intra-ensemble spread** (STD in member dimension averaged across start dates)

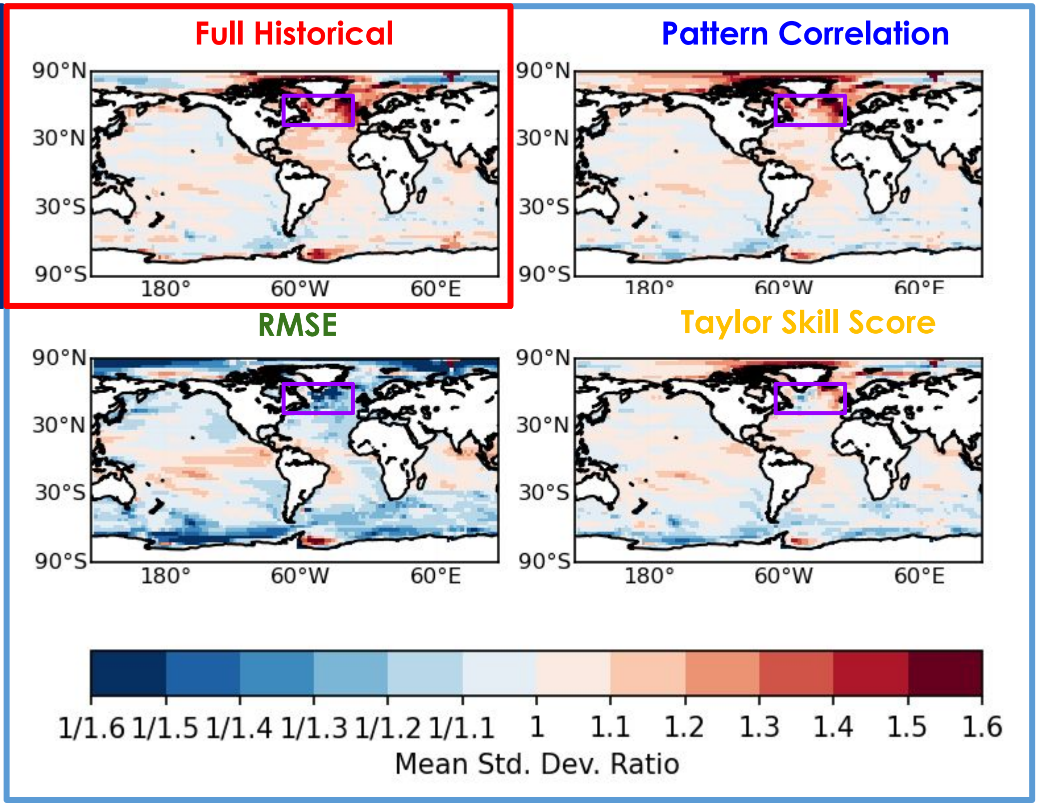


# IMPACT OF THE SELECTED APPROACH ON SPREAD

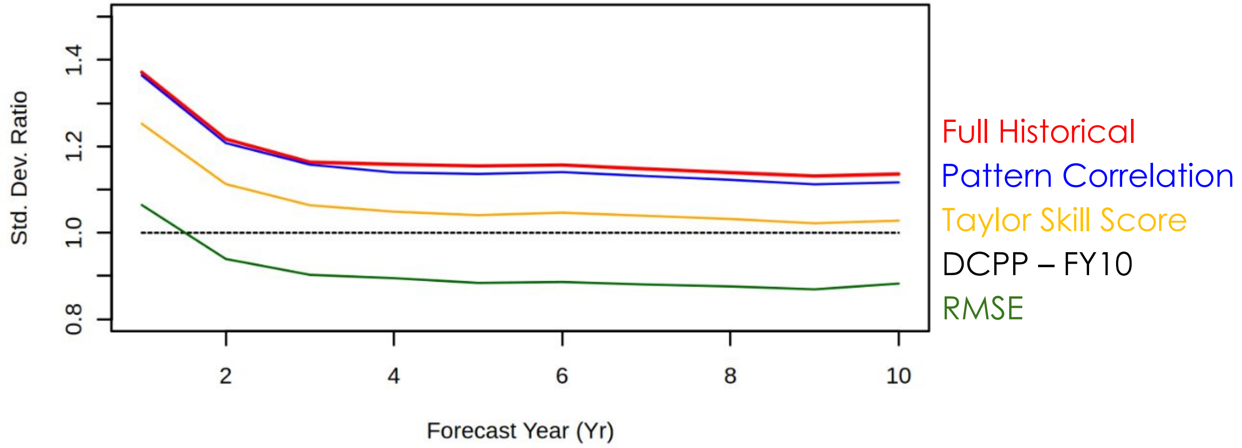
**Mean intra-ensemble spread** (STD in member dimension averaged across start dates)

**Target spread** STD ratio vs DCPP-FY10 **(Global Domain)**




Spread ratio in **Subpolar North Atlantic**

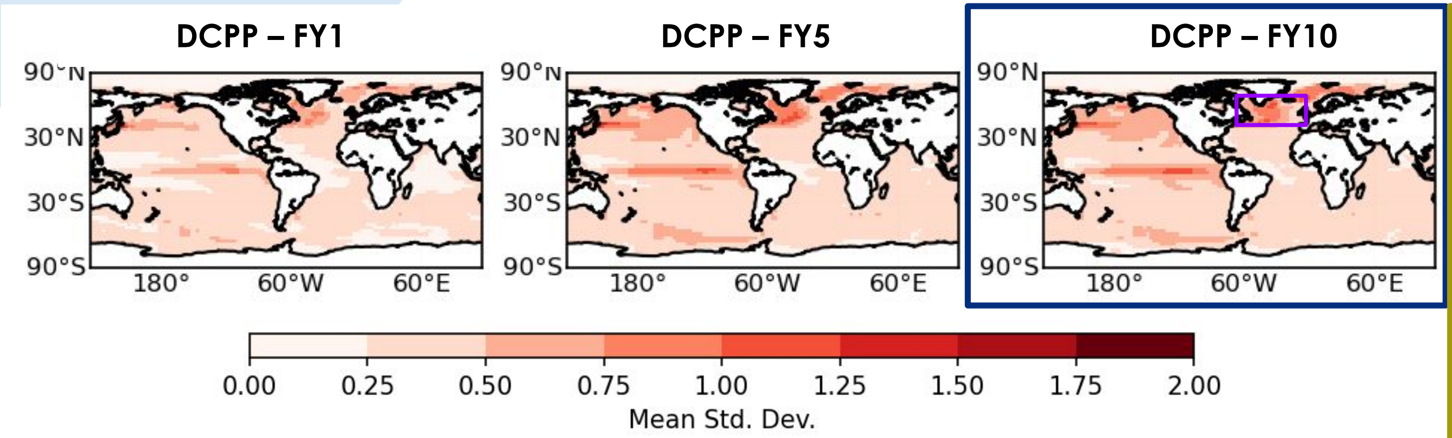


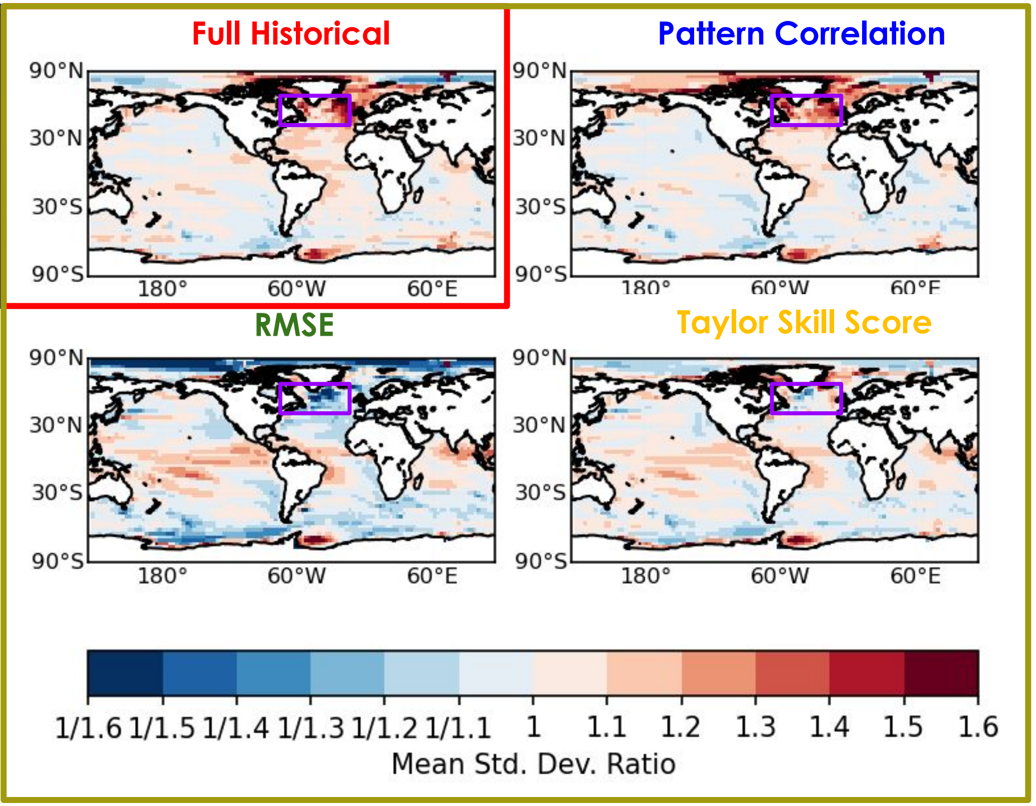
# IMPACT OF THE SELECTED APPROACH ON SPREAD

**Mean intra-ensemble spread** (STD in member dimension averaged across start dates)

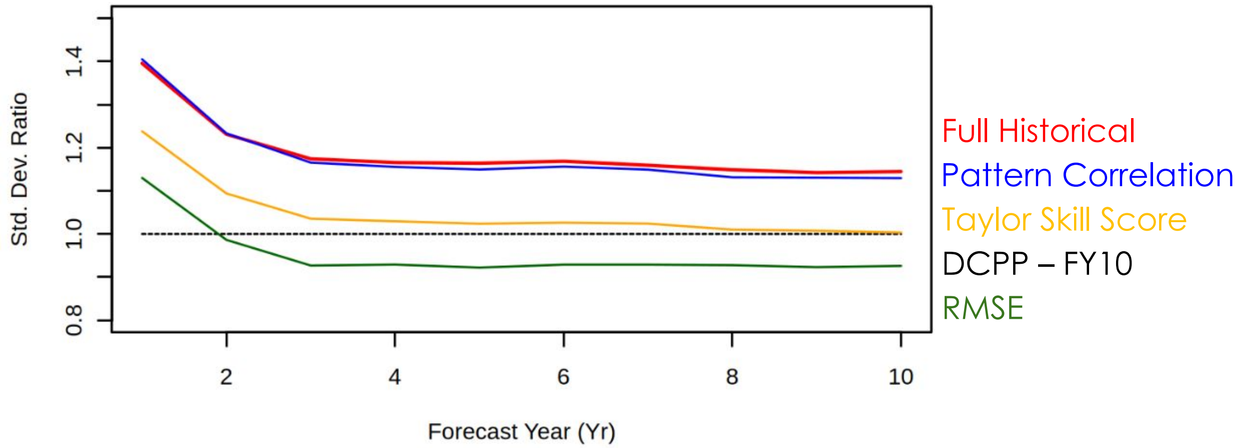
**Target spread**

STD ratio vs DCPP-FY10 **(North Atlantic)**



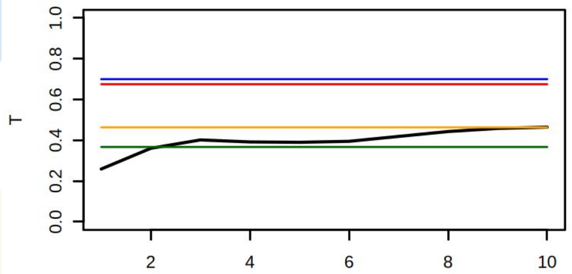

Spread ratio in **Subpolar North Atlantic**



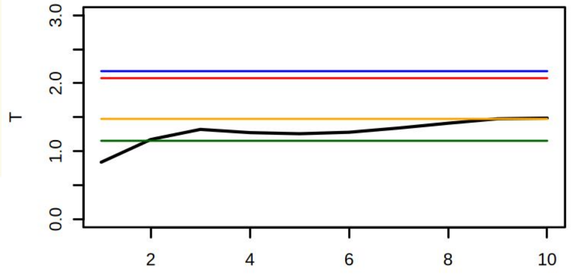
# IMPACT OF THE SELECTED APPROACH ON SPREAD

## Spread for other percentiles (North Atlantic domain)

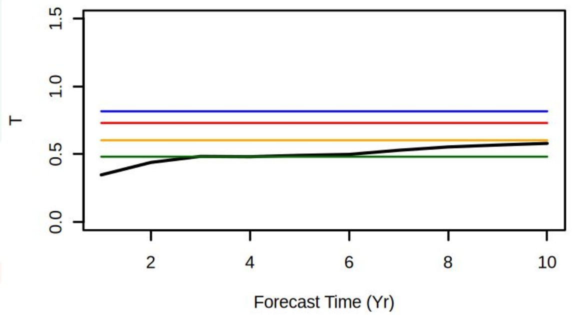
Mean **STD across members**



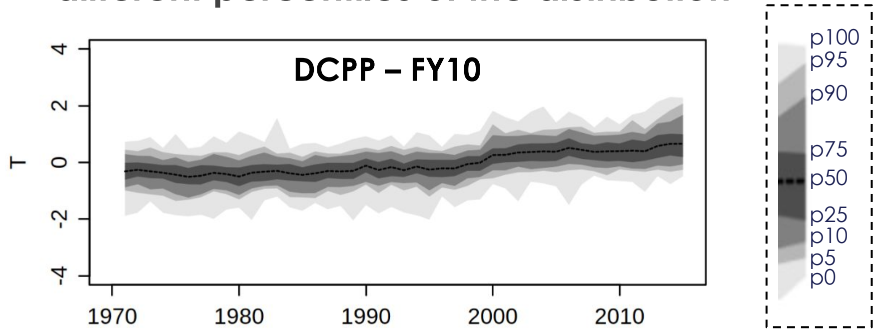

Mean **p95-p5 difference**

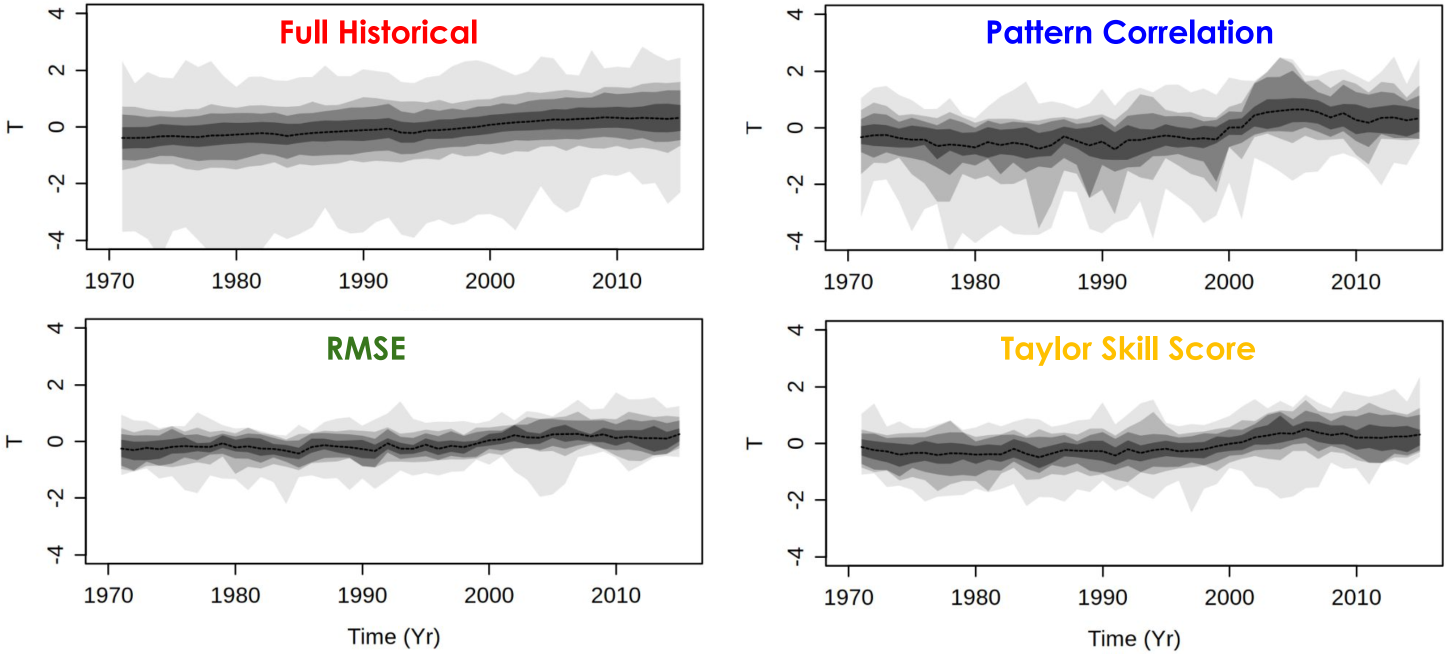



Mean **Inter-Quartile Range**




North Atlantic SST **evolution for different percentiles of the distribution**

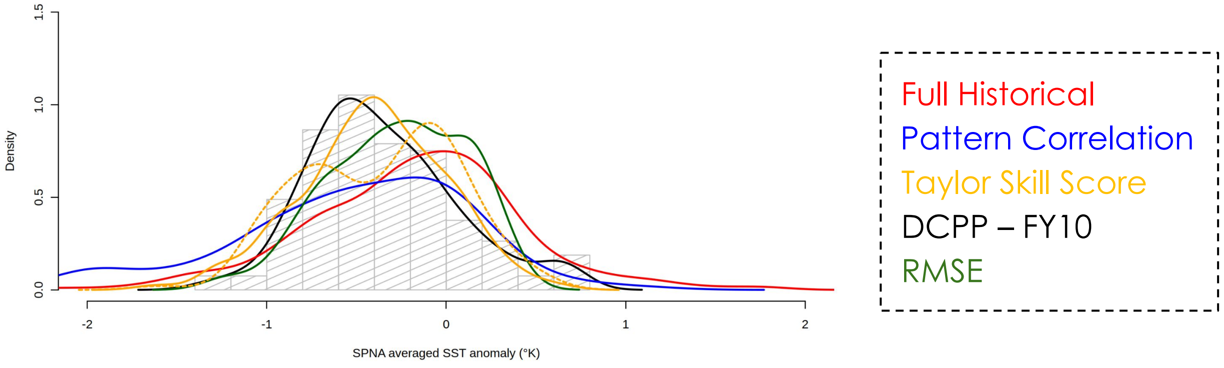


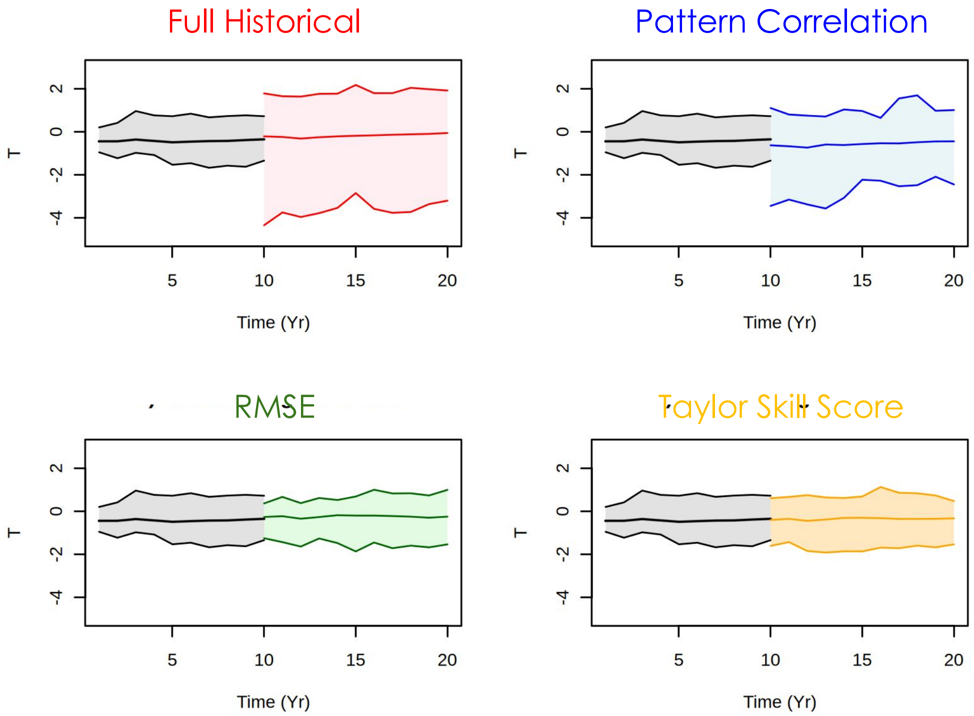
# TESTS ON DISTRIBUTION CONSISTENCY

## Ensemble **distribution** (North Atlantic domain)

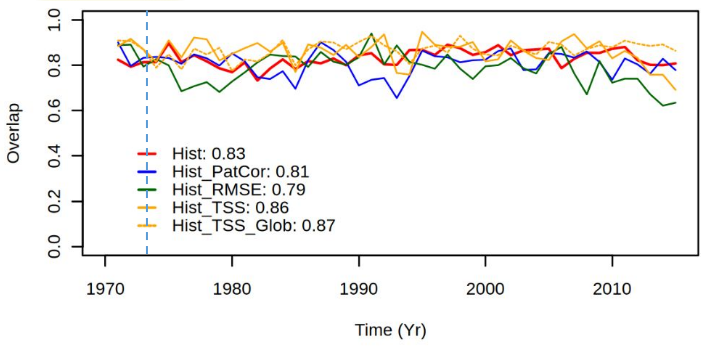
### **PDF** of SPNA tos for example start date 1973



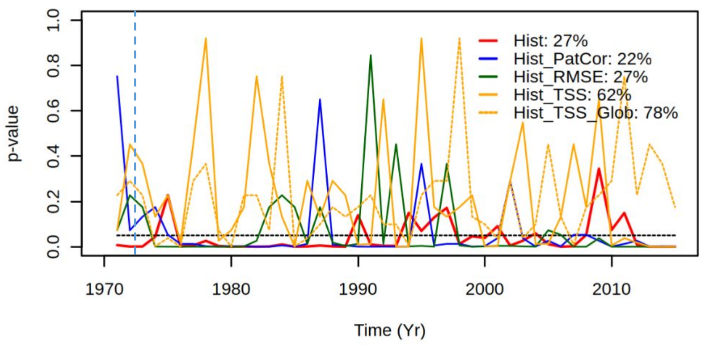

### **Full spread consistency** for 1973 start date




### **Overlap** with DCPP - FY10




### **P-value KS test** wrt DCPP - FY10



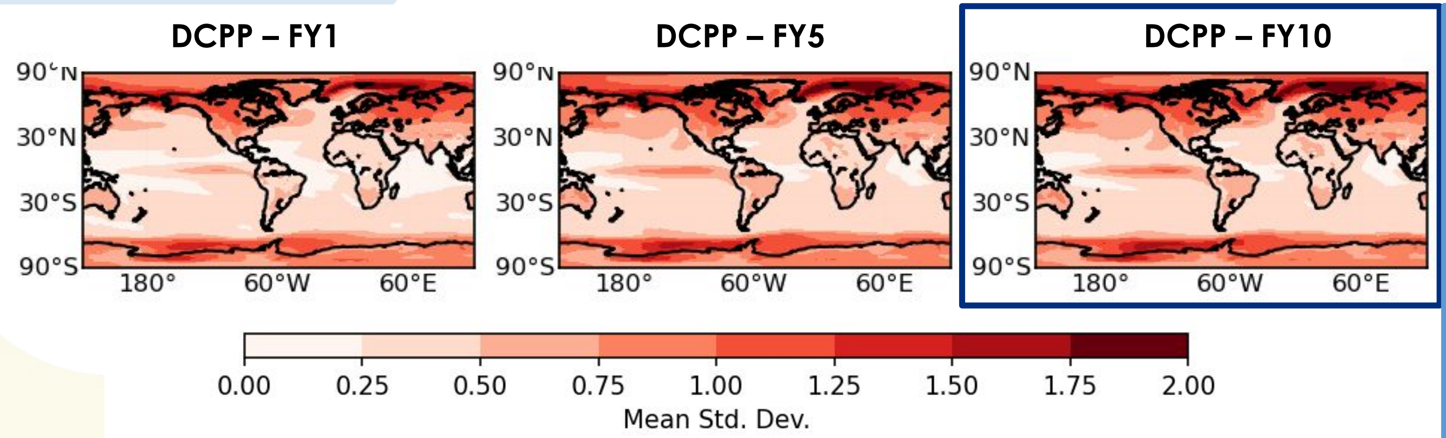
# EFFECTIVENESS OF THE METHOD IN OTHER VARIABLES

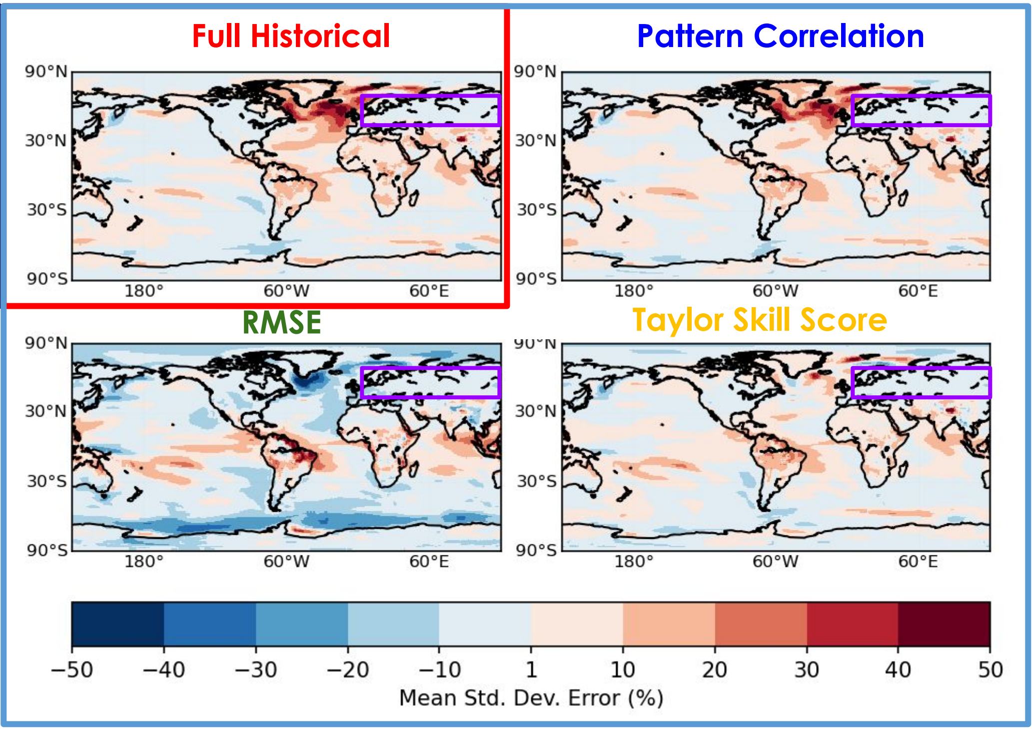
Spread comparison for **surface air temperature**

**Target spread**

STD ratio vs DCPP-FY10

**(Global domain)**

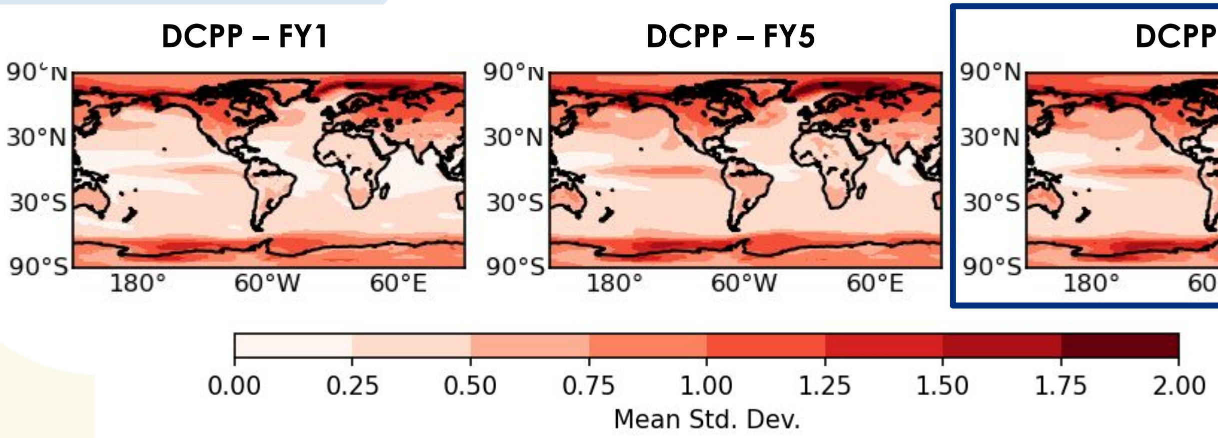


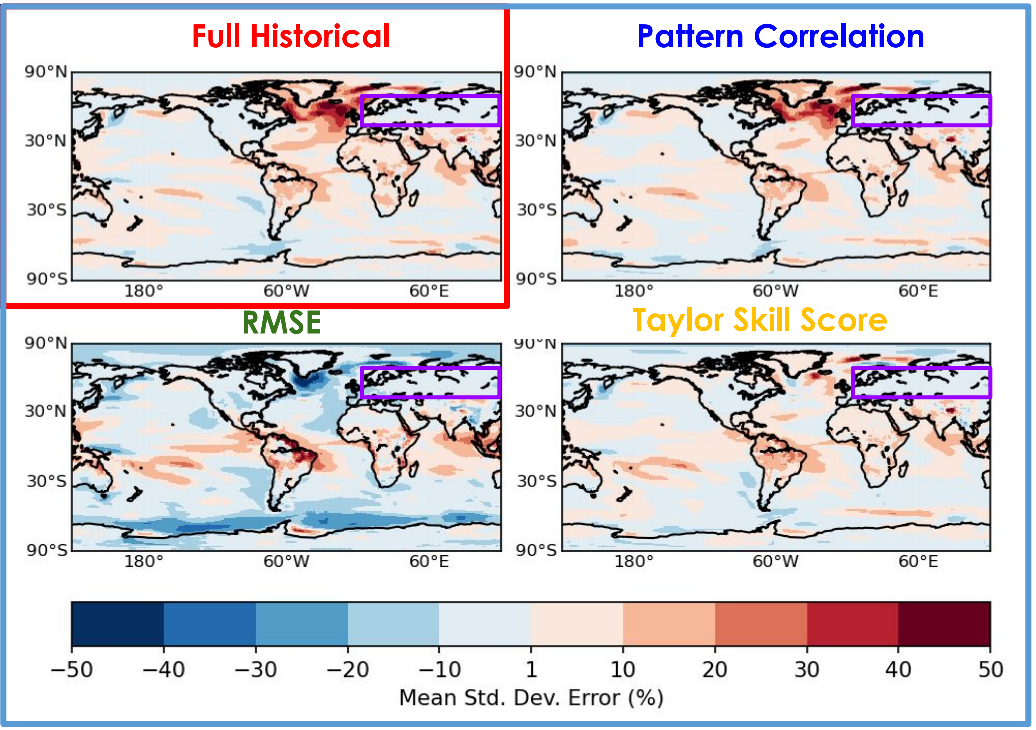
# EFFECTIVENESS OF THE METHOD IN OTHER VARIABLES

Spread comparison for **surface air temperature**

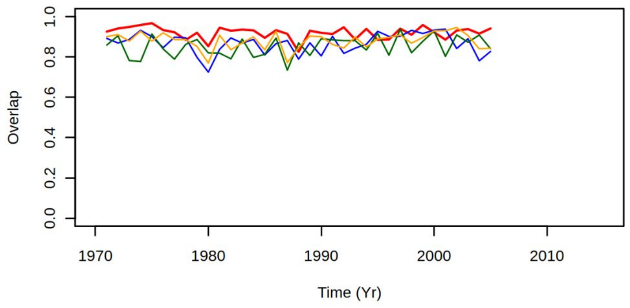
**Target spread**



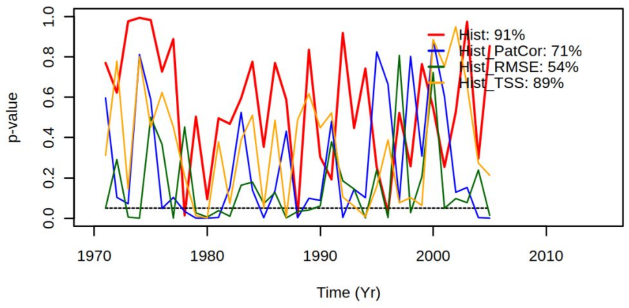

STD ratio vs DCPP-FY10 (Global domain)




**Overlap** with DCPP - FY10




**P-value KS test** wrt DCPP - FY10



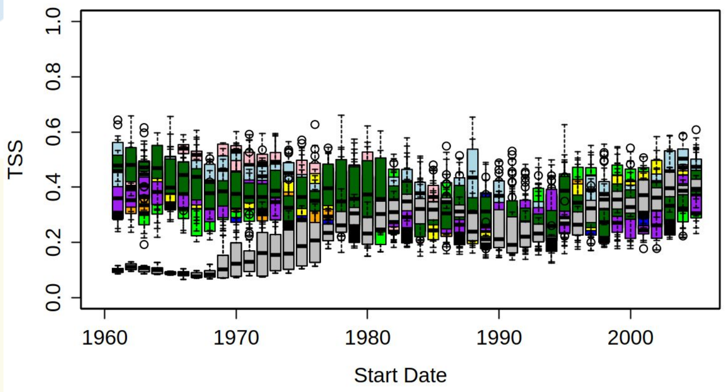
# TAKE HOME MESSAGES

- We present a new method to blend climate projections and decadal predictions that **ensures physical consistency** by identifying **analogs between predicted and projected SST-anomaly patterns**.
- **Statistical consistency** is achieved by selecting **analogs separately for each member** of the decadal prediction ensemble.
- We introduce the **Taylor Skill Score as a new metric for analog selection**, which substantially **improves the statistical agreement** between the blended products.
- The blending **performs well for SST,** particularly **in the North Atlantic**, **where inconsistencies between projections and predictions are found to be largest**.
- However, the method **does not fully resolve inconsistencies in some regions for atmospheric air temperature**, for which **additional constraints** – such as matching large-scale atmospheric circulation patterns – **may be required**.

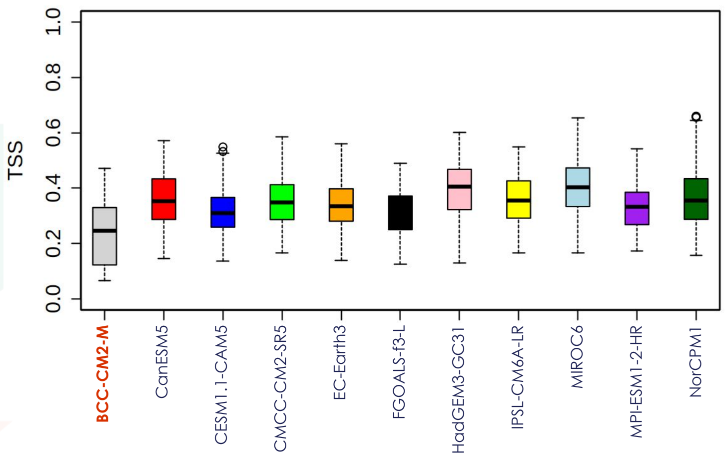
# TEMPORAL CONSISTENCY OF ANALOGS PER DPS

## Global domain (Centered Taylor Skill Score)

TSS for highest ranked member (per start date)

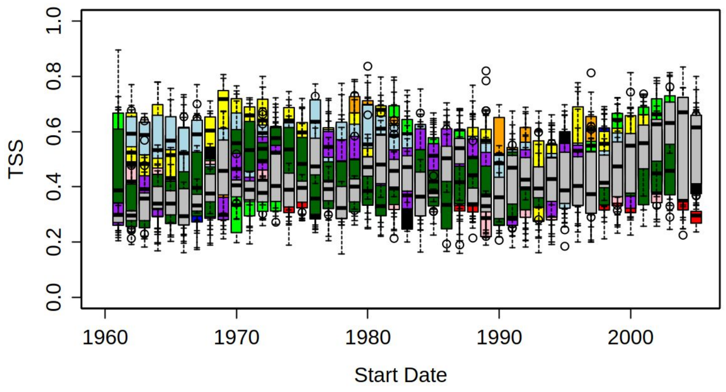



TSS for highest ranked member (all start dates)

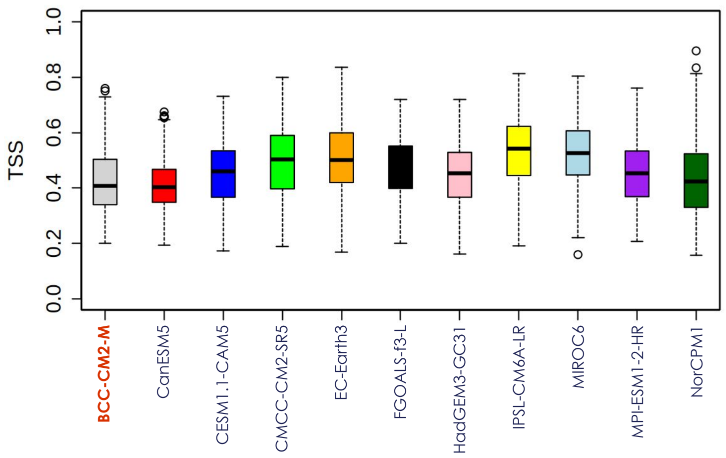



## North Atlantic domain (Centered Taylor Skill Score)

TSS for highest ranked member (per start date)



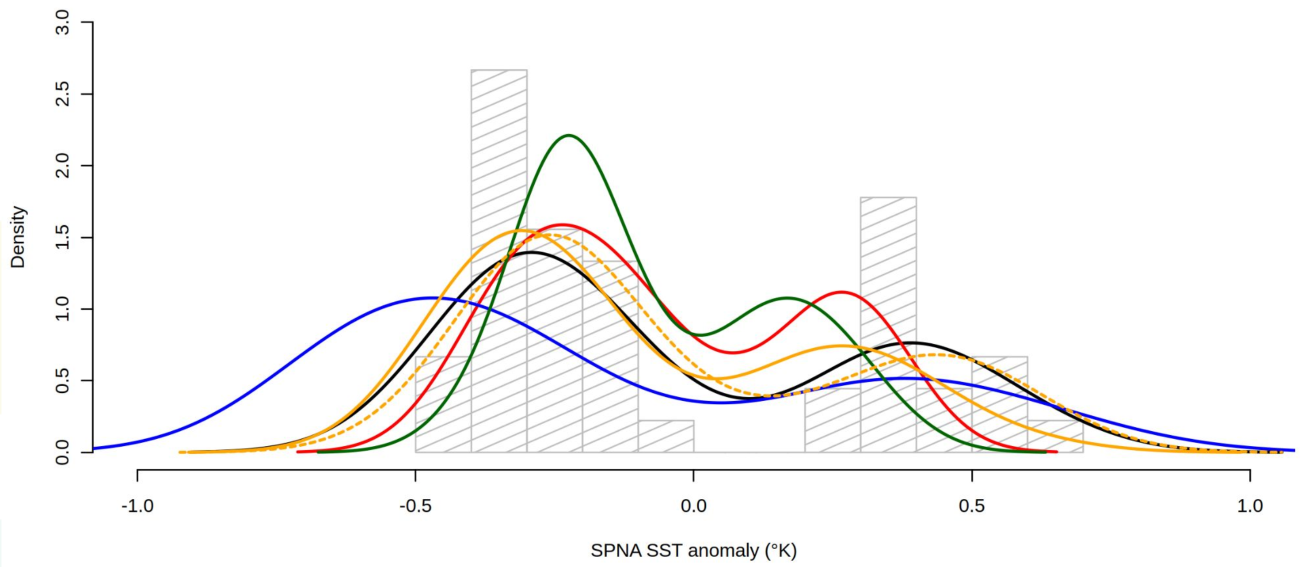

TSS for highest ranked member (all start dates)



# TESTS ON DISTRIBUTION CONSISTENCY

## Ensemble **mean** **(North Atlantic domain)**

**PDF** of SPNA tos ensemble means (per start date)



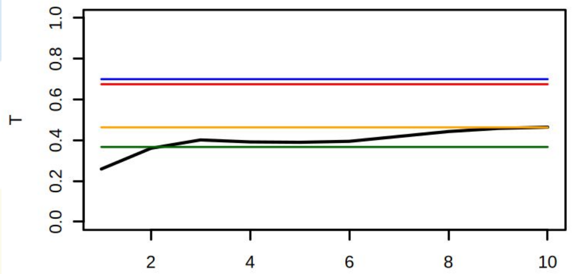

| **Ensemble** | **Overlap** | **KS test p-value** | **Ensemble STD** |
|---|---|---|---|
| DCPP – FY10 | 1.000 | 1.000 | 0.348 |
| Full Historical NA | 0.769 | 0.081 | 0.251 |
| Pattern Correlation NA | 0.750 | 0.013 | 0.438 |
| RMSE NA | 0.670 | 0.046 | 0.212 |
| Taylor Skill Score NA | 0.900 | 0.653 | 0.306 |
| **Taylor Skill Score Global** | **0.934** | **0.824** | **0.338** |

The largest DCPP-Historical spread discrepancies in SST occur in the SPNA, for which the Taylor Skill Score metric computed over the North Atlantic domain provides the most consistent results with DCPP

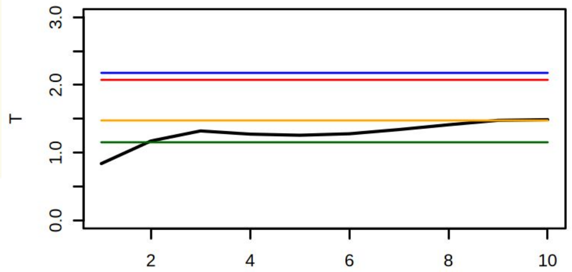
# IMPACT OF THE SELECTED APPROACH ON SPREAD

## Spread for other percentiles (North Atlantic domain)

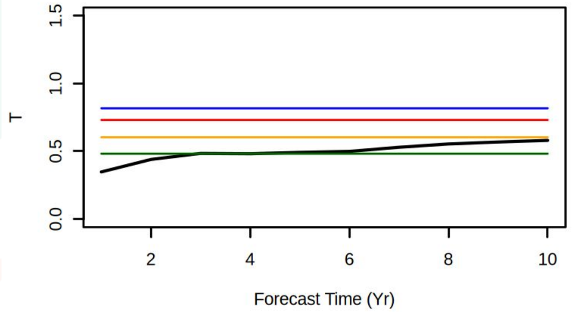
Mean **STD across members**




Mean **p95-p5 difference**

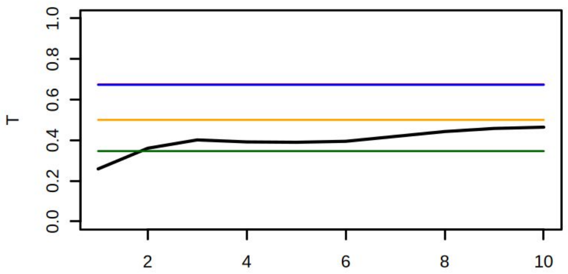



Mean **Inter-Quartile Range**

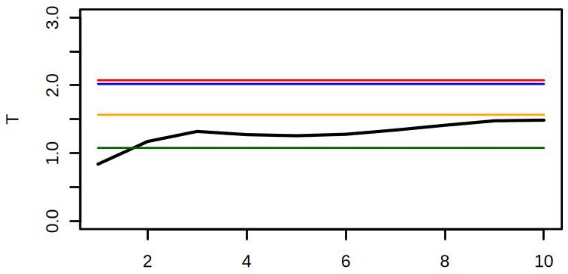



## (Global domain)

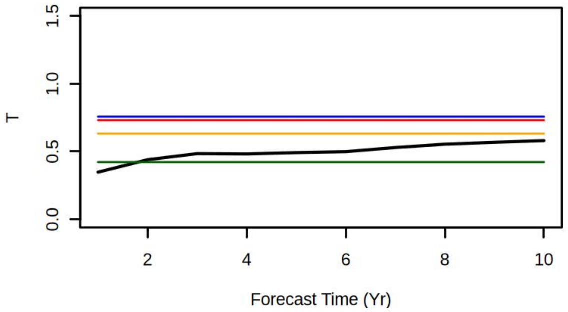
Mean **STD across members**




Mean **p95-p5 difference**




Mean **Inter-Quartile Range**



# TAYLOR SKILL SCORE

Term that assesses **if the forecast pattern is correct**

$$\mathrm{TSS} = \frac{\boxed{2(1+R)}}{\boxed{(\sigma_f/\sigma_o + \sigma_o/\sigma_f)^2}}$$

Term that assesses **if the forecast simulates the right amplitude**

Where:

- R = correlation between forecast and observations
- $\sigma_f$ = standard deviation of the forecast
- $\sigma_o$ = standard deviation of the observations

# TAYLOR SKILL SCORE*

$$\mathrm{TSS} = \frac{(1+R)^4}{4\left(\frac{\sigma_f}{\sigma_o} + \frac{\sigma_o}{\sigma_f}\right)^2}$$

Where:

- R = correlation between forecast and observations
- $\sigma_f$ = standard deviation of the forecast
- $\sigma_o$ = standard deviation of the observations

The **Taylor Skill Score (TSS)** is a metric used to evaluate how well a model reproduces observed data. It condenses three key statistical comparisons into a single number:

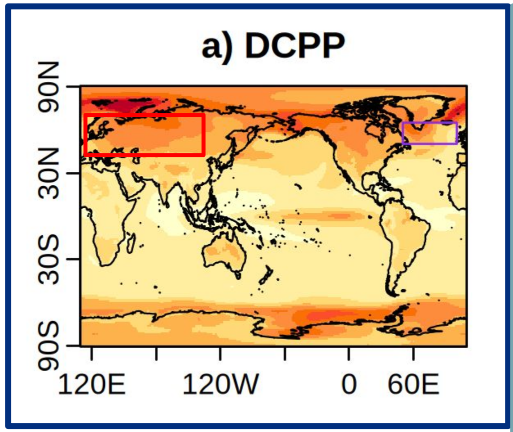
1. **Correlation** (how well the model captures the *pattern* of variability)
2. **Standard deviation ratio** (how well the model captures the *amplitude* of variability)
3. **Centered RMSE** (how closely the model matches observations *after removing the mean bias*)

***this version penalizes low correlation values***

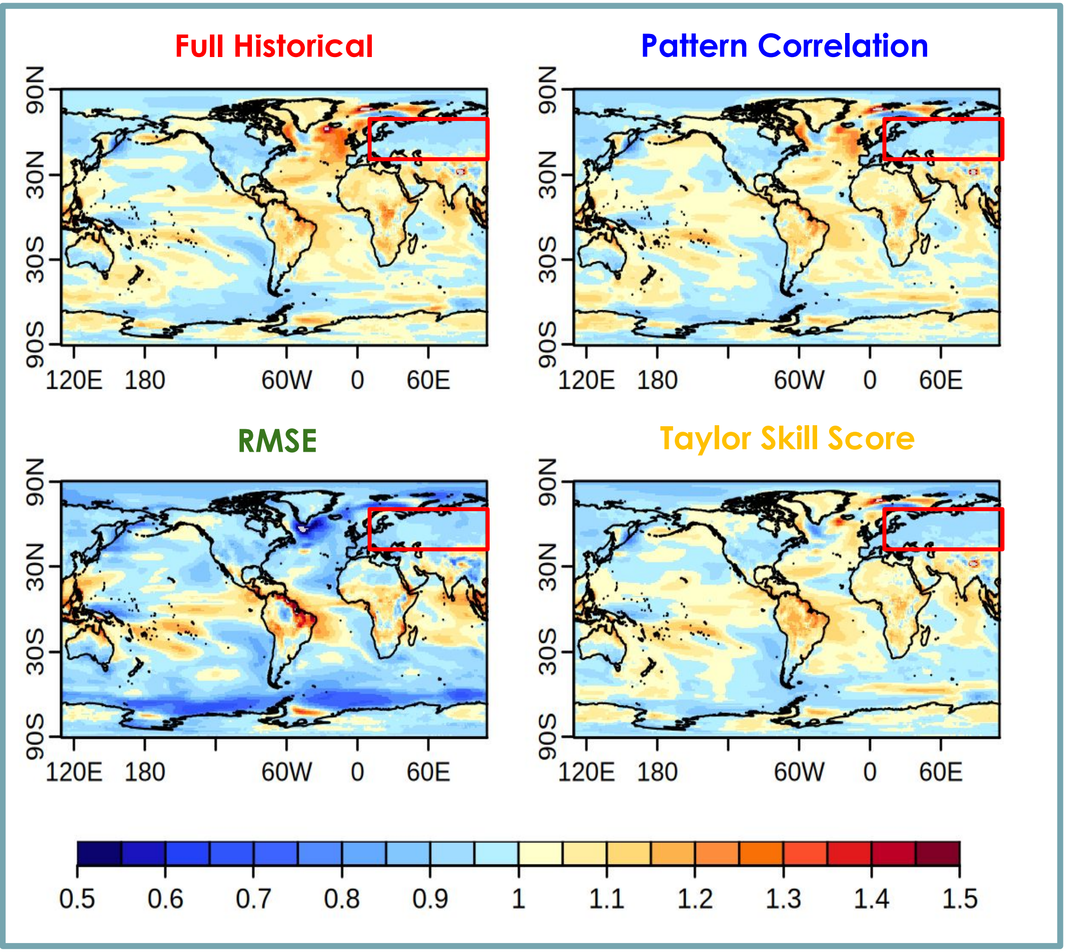
# EFFECTIVENESS OF THE BLENDING IN OTHER DOMAINS

Spread comparison for **surface air temperature**

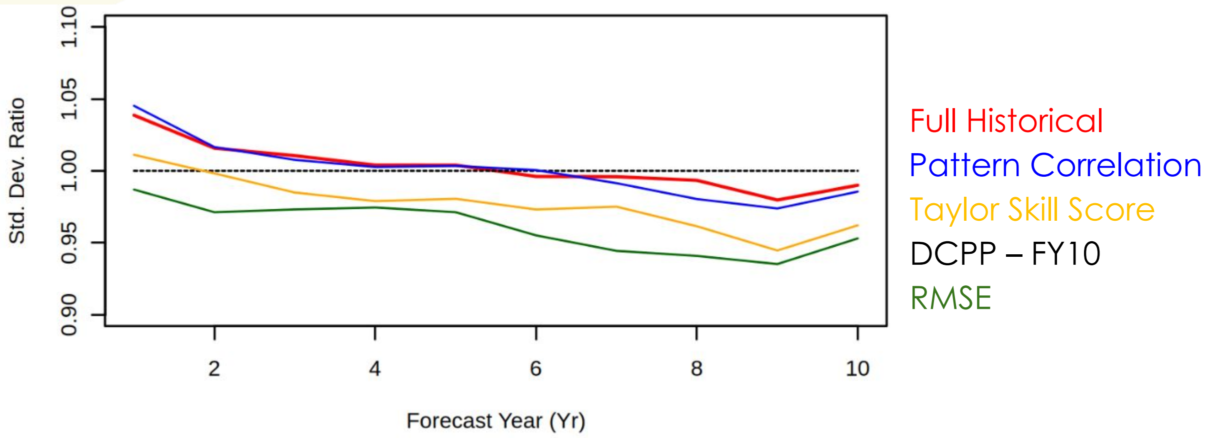
**Target spread**




**North Atlantic** – Spread ratio vs DCPP-FY10

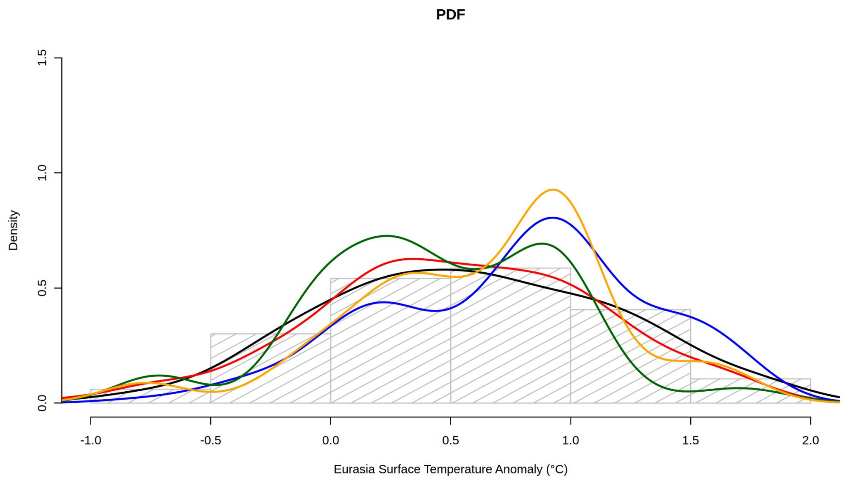



Spread ratio in **Eurasia**

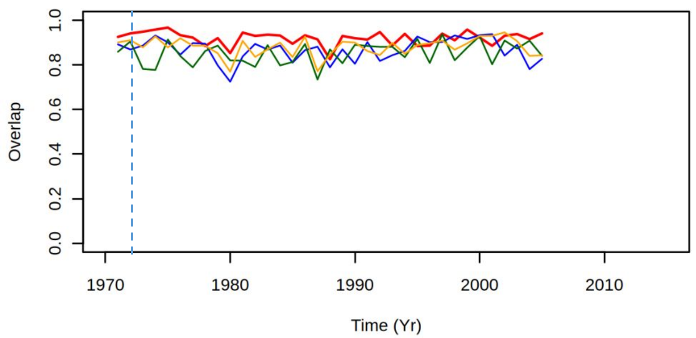


# EFFECTIVENESS OF THE BLENDING IN OTHER DOMAINS

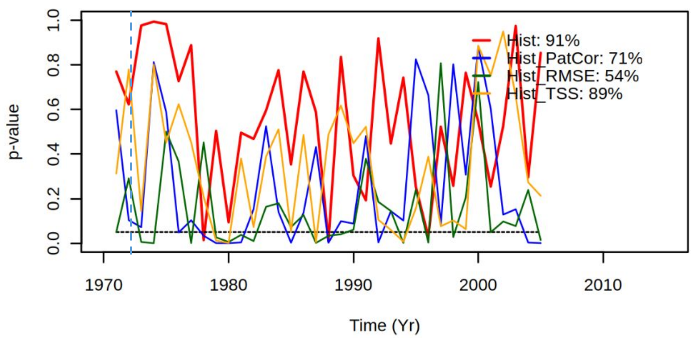
Ensemble **distribution** (North Atlantic domain)




**Overlap** with DCPP - FY10



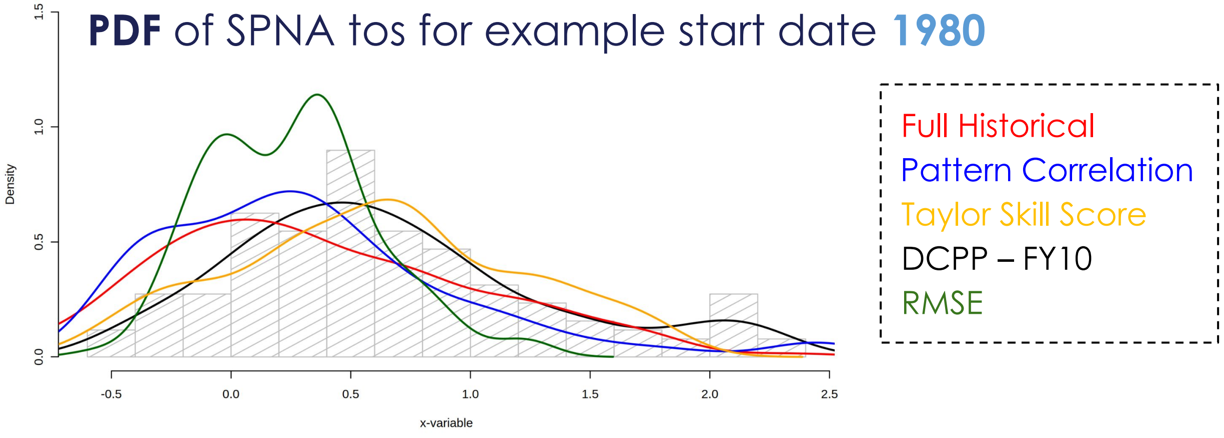

**P-value KS test** wrt DCPP - FY10



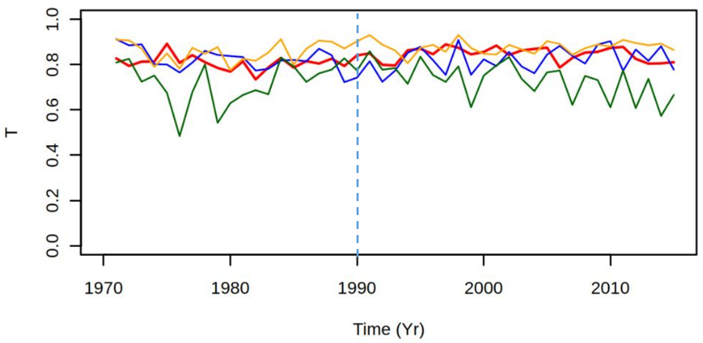
# TESTS ON DISTRIBUTION CONSISTENCY

Ensemble **distribution** (Global domain)

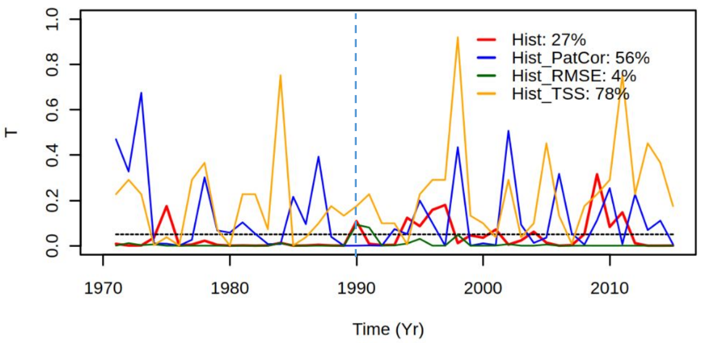
**PDF** of SPNA tos for example start date **1980**




**Overlap** with DCPP - FY10



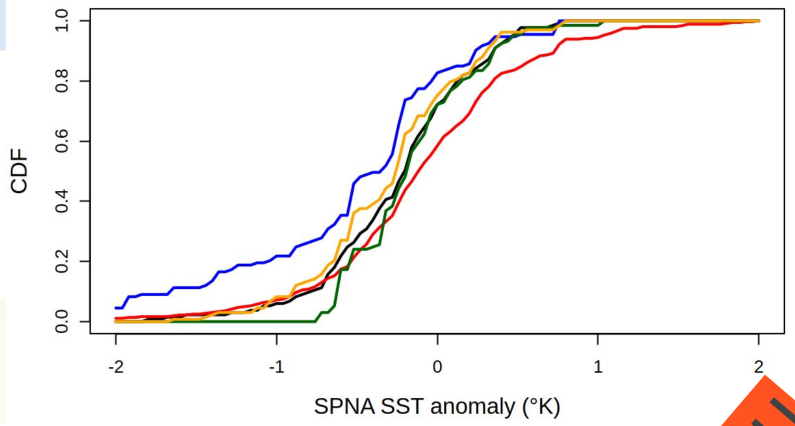

**P-value KS test** wrt DCPP - FY10



# TESTS ON DISTRIBUTION CONSISTENCY

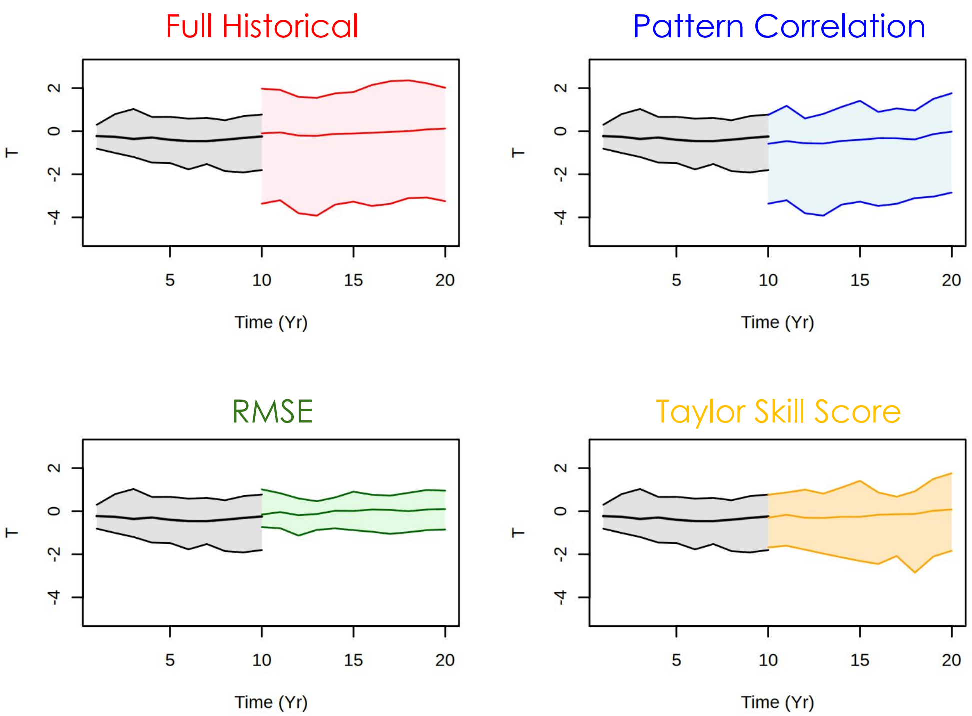
Ensemble **distribution** (North Atlantic domain)

**CDF** of SPNA tos for example start date 1980

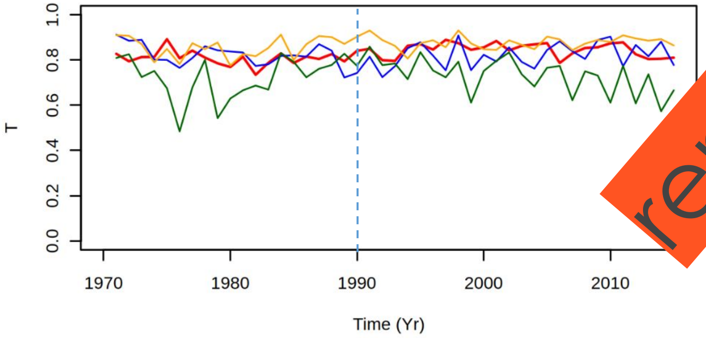



Full Historical
Pattern Correlation
Taylor Skill Score
DCPP – FY10
RMSE

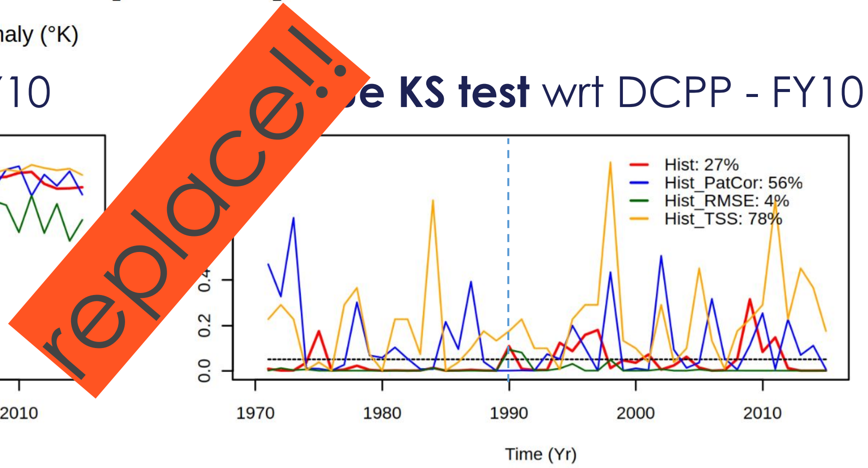
**Full spread consistency** for 1980 start date




**Overlap** with DCPP - FY10



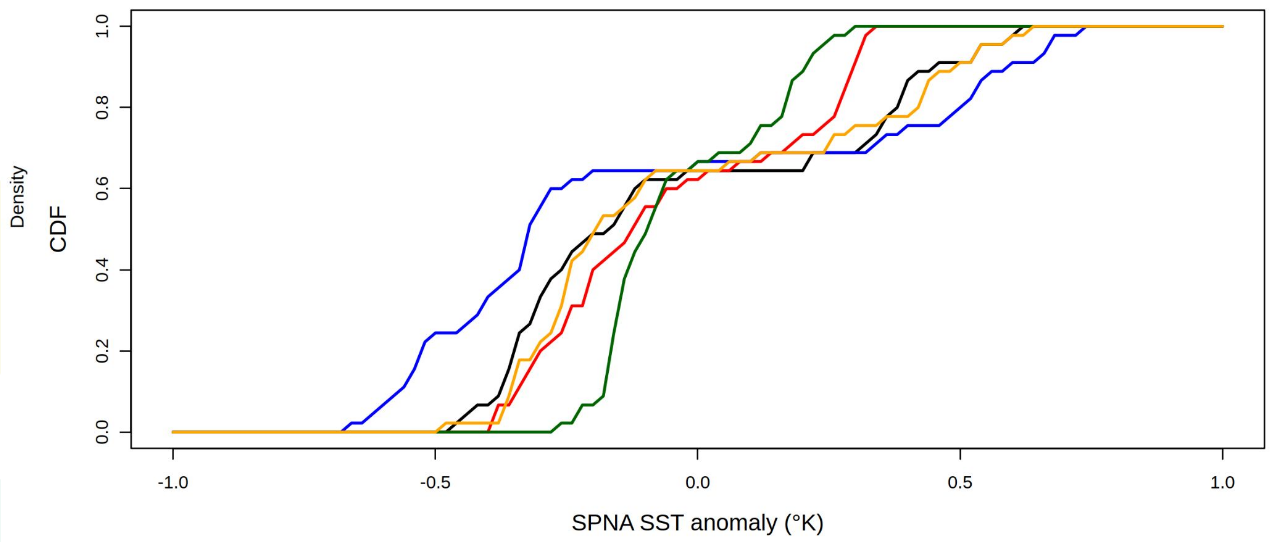

**e KS test** wrt DCPP - FY10



# TESTS ON DISTRIBUTION CONSISTENCY

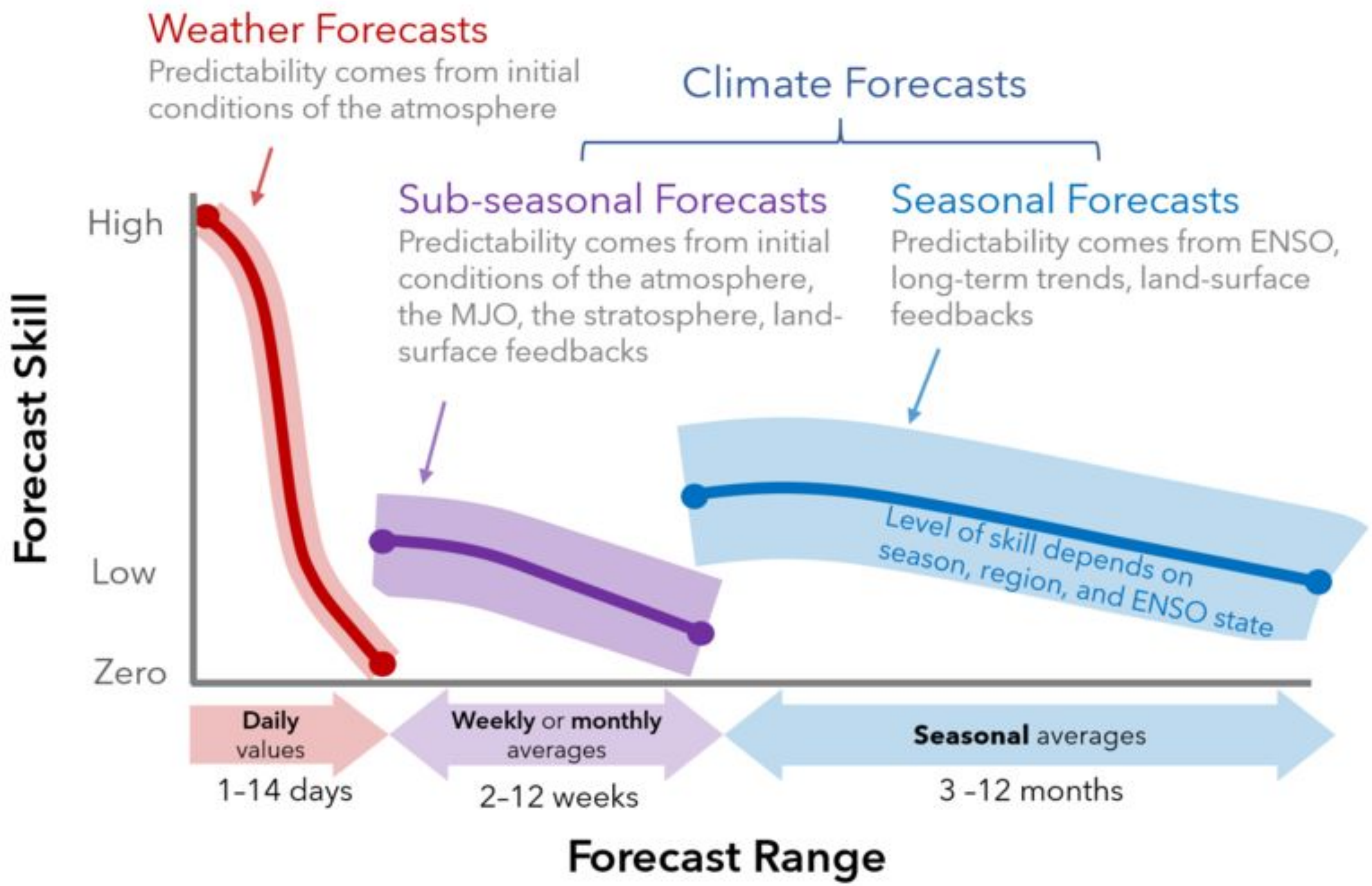
## Ensemble **mean** (North Atlantic domain)

**CDF** of SPNA tos ensemble means (per start date)




| Ensemble | Overlap | KS test p-value | Ensemble STD |
|---|---|---|---|
| DCPP – FY10 | 1.000 | 1.000 | 0.348 |
| Full Historical | 0.769 | 0.081 | 0.251 |
| Pattern Correlation | 0.750 | 0.046 | 0.459 |
| RMSE | 0.476 | 0.001 | 0.160 |
| Taylor Skill Score | **0.934** | **0.824** | **0.338** |

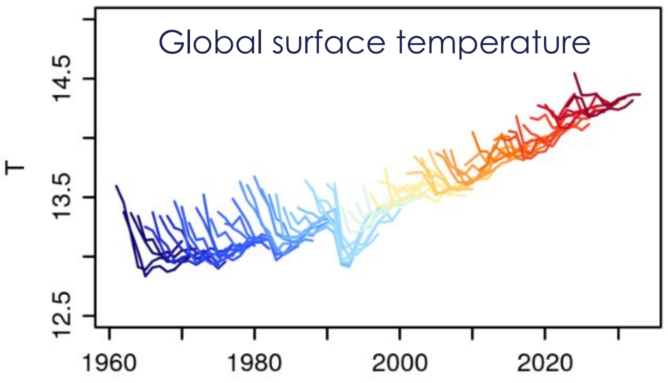
The largest DCPP-Historical spread discrepancies in SST occur in the SPNA, for which the Taylor Skill Score metric computed over the North Atlantic domain provides the most consistent results with DCPP


Weather Forecasts
Predictability comes from initial conditions of the atmosphere
Climate Forecasts
Sub-seasonal Forecasts
Predictability comes from initial conditions of the atmosphere, the MJO, the stratosphere, land-surface feedbacks
Seasonal Forecasts
Predictability comes from ENSO, long-term trends, land-surface feedbacks
Level of skill depends on season, region, and ENSO state
Forecast Skill
High
Low
Zero
Daily values
1-14 days
Weekly or monthly averages
2-12 weeks
Seasonal averages
3 -12 months
Forecast Range

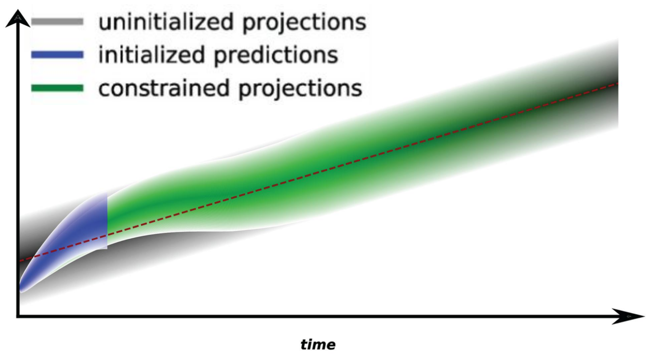
# REVEALING INCONSISTENCIES

| PREDICTIONS (decadal) | PROJECTIONS |
|---|---|
| Initialized from **observed state** | Based on **uninitialized** ensembles |
| Subject to important **model drifts** | Model is already in its **preferred state** |
| *Ensemble mean* captures **internal+forced predictability** | *Ensemble mean* only captures **forced predictability** |
| *Ensemble spread* **constrained by the initial state** | *Ensemble spread* seeks to **include all internal variability phases** |
| Retrospective predictions required with **very high computational costs** | **Continuous simulations** (only a few members needed) |
| **Two temporal dimensions** (forecast time and initialization time) | **Only one temporal dimension** (simulated year) |
| Only about **a dozen modelling centers** worldwide produce them | **All major modelling centers (>40)** worldwide produce them |

Drift Illustration
*Courtesy of R. Bilbao*



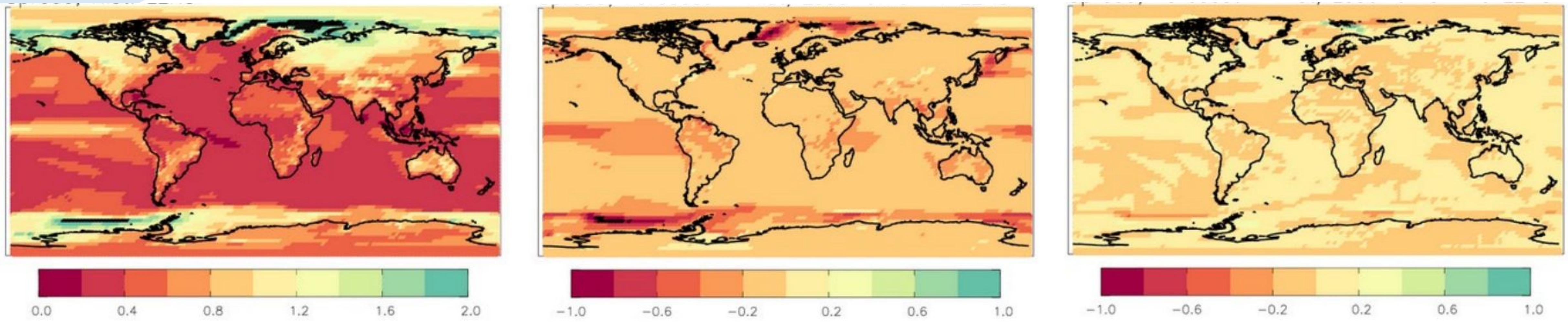

Ensemble Comparison
*Befort et al.(2020)*



# REVEALING INCONSISTENCIES: Ensemble Spread

Mean inter-model ensemble spread of surface air temperature

**Historical** | **Prediction (FY1) - Historical** | **Prediction (FY10) - Historical**

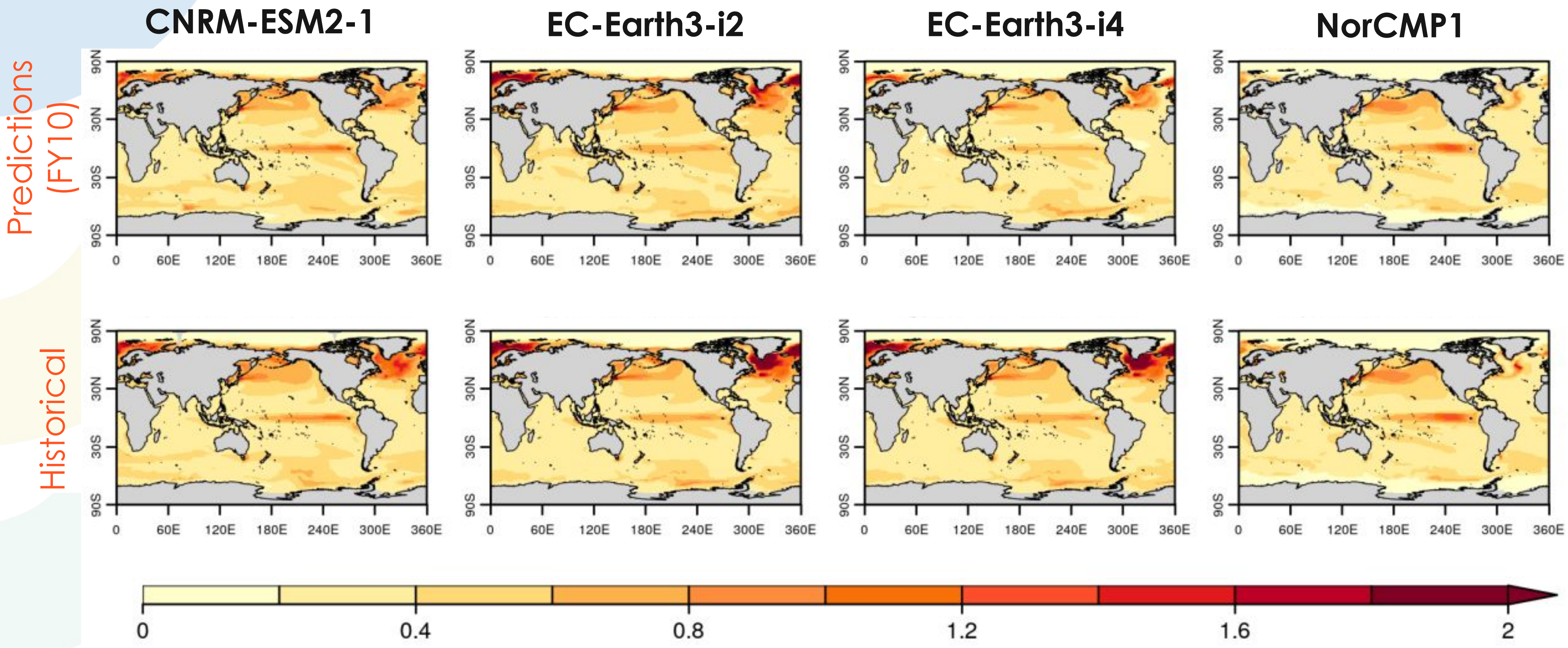



**NCAR Large Ensembles**

***DMI:*** *Christensen et al (2023)*

# REVEALING INCONSISTENCIES: Ensemble Spread

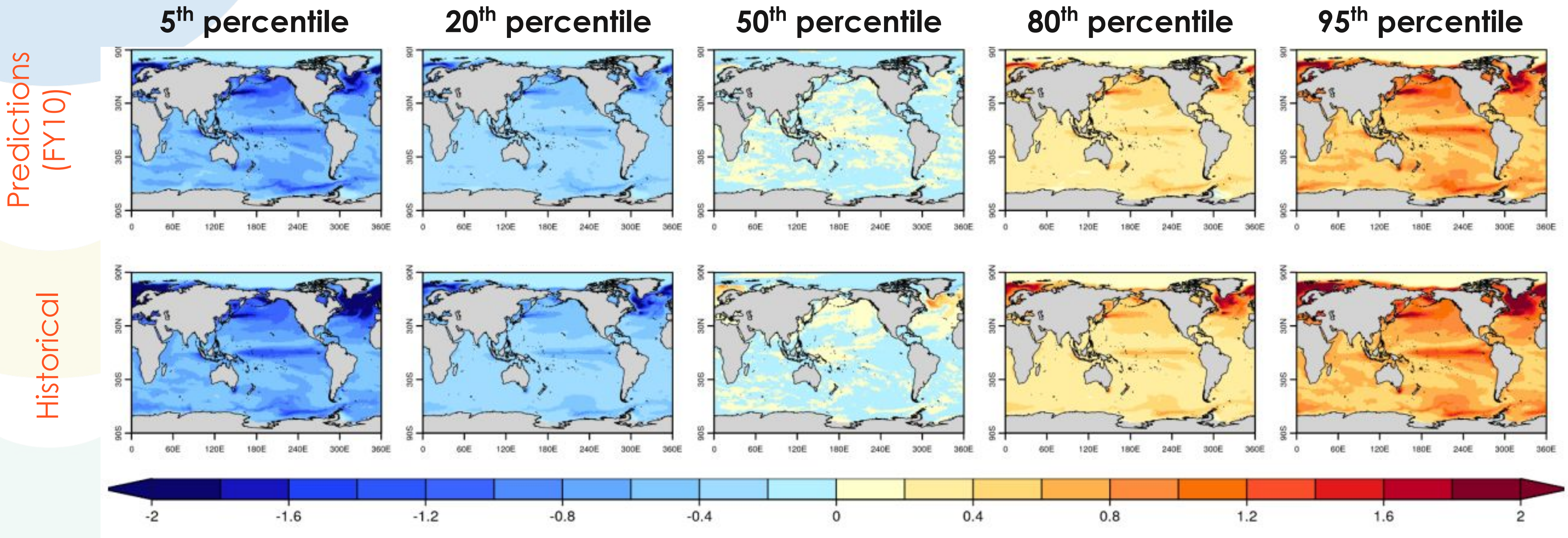
Mean inter-model ensemble spread of sea surface temperature




***BSC:** R. Bilbao and co-authors*

# REVEALING INCONSISTENCIES: Ensemble Spread

Selected percentiles in the inter-model ensemble of sea surface temperature

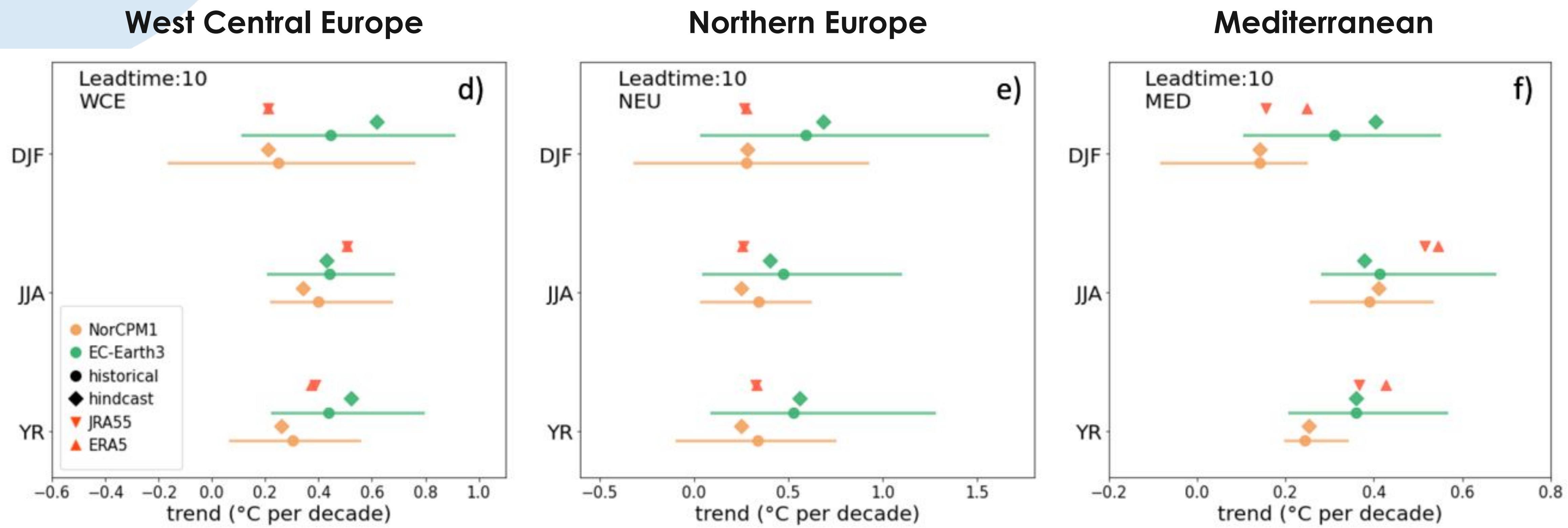



For **EC-Earth3** (i2)

***BSC:*** *R. Bilbao and co-authors*

# REVEALING INCONSISTENCIES: Long-term trends

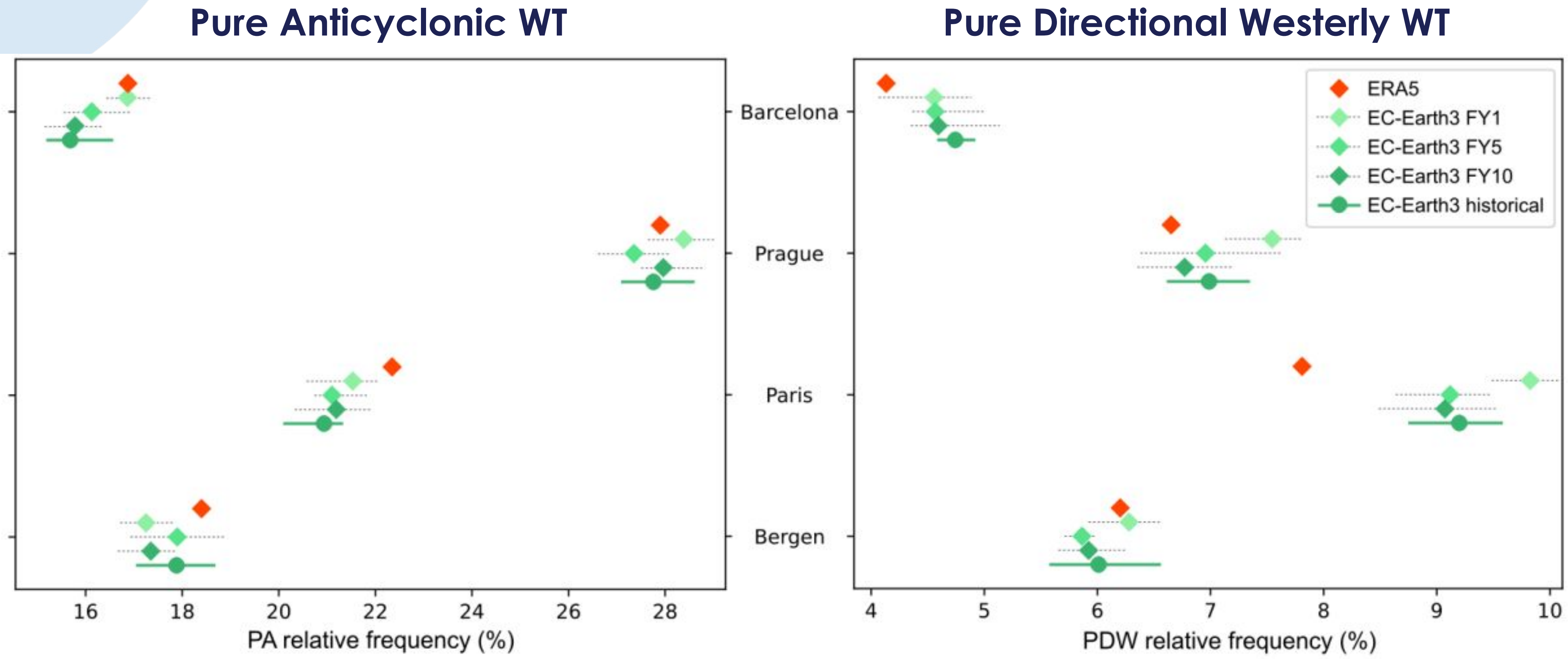
Ensemble mean regional surface temperature trends




***CNRS-CERFACS:*** *R. Bonnet and co-authors*

# REVEALING INCONSISTENCIES: Dynamical drivers

Mean relative frequency of main weather types in the four I4C demonstrator cities

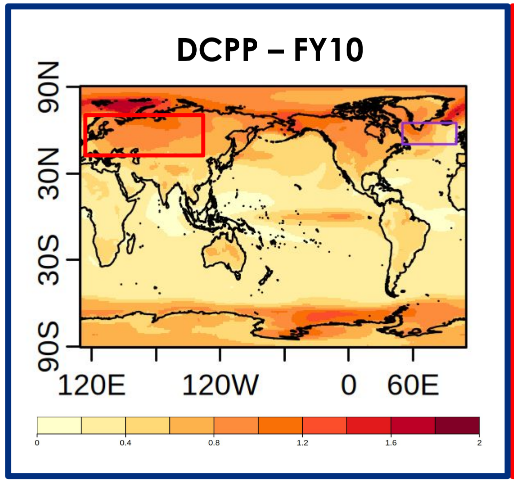


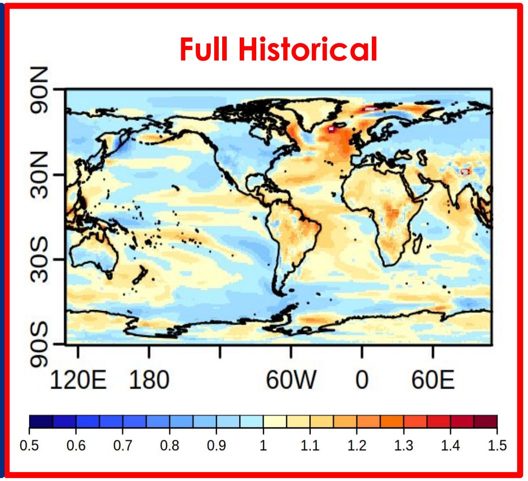

***CSIC:*** *S. Brands and co-authors*

# EFFECTIVENESS OF THE METHOD IN OTHER VARIABLES

Spread comparison for **surface air temperature**

**Target spread** Spread ratio vs DCPP-FY10

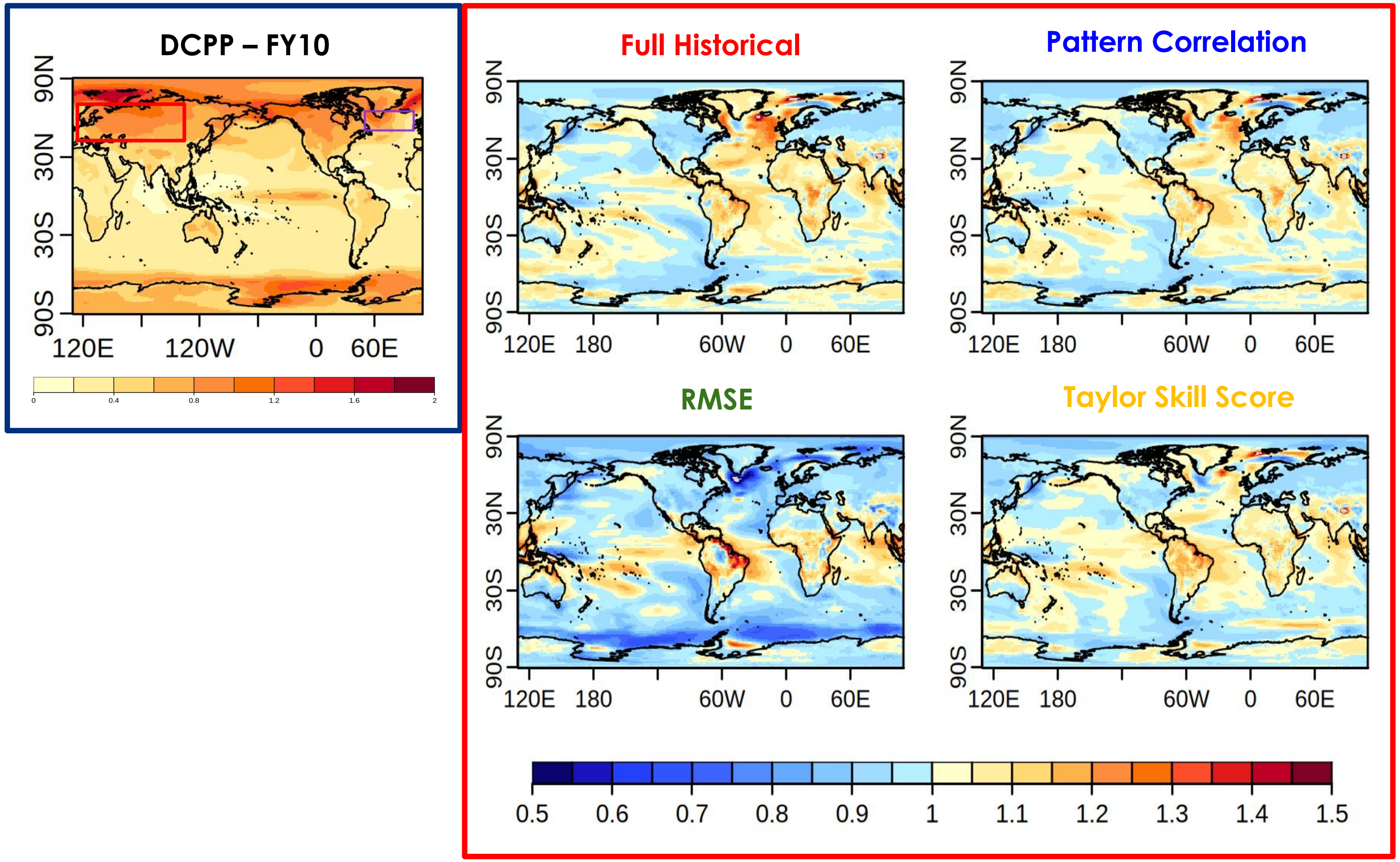


# EFFECTIVENESS OF THE METHOD IN OTHER VARIABLES

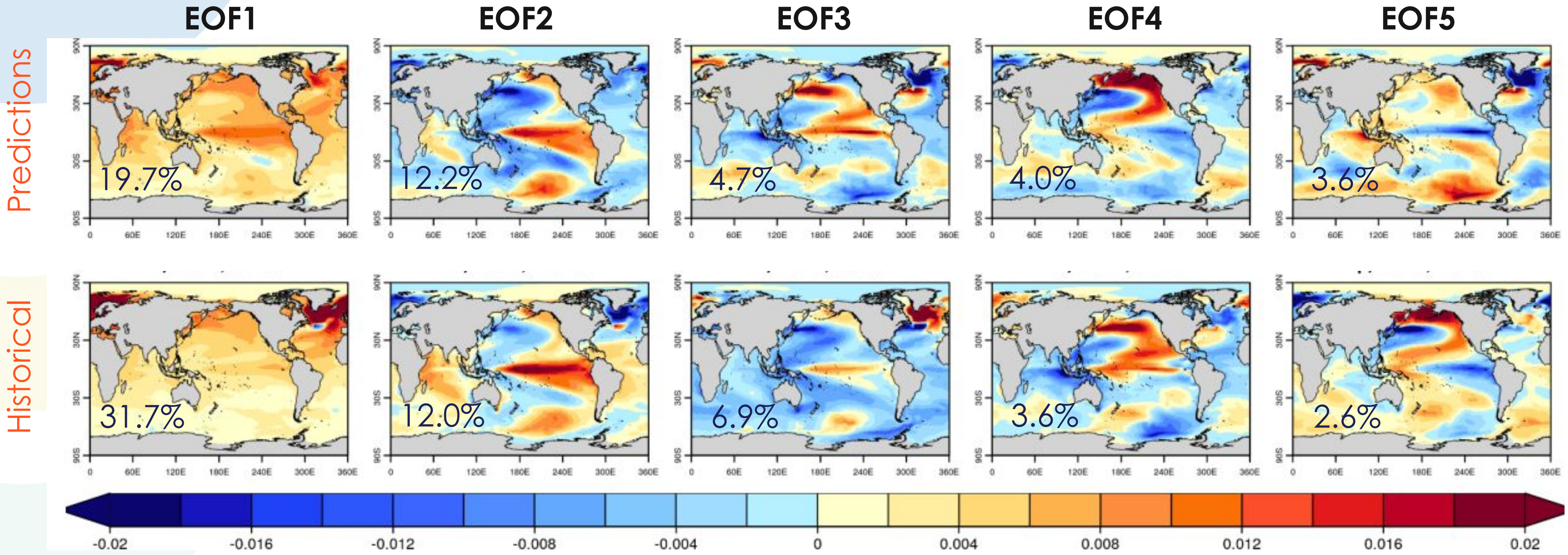
Spread comparison for **surface air temperature**

**Target spread** | **Global** – Spread ratio vs DCPP-FY10



# REVEALING INCONSISTENCIES: Large-scale drivers

Leading EOFs of the SST anomaly fields (in the member dimension)




For **EC-Earth3** (i2)

***BSC:*** *R. Bilbao and co-authors*