Local versus non-local calibration techniques

Ángel G. Muñoz

with contributions from
Simon J. Mason
Andrew W. Robertson

International Research Institute for Climate and Society
Earth Institute | Columbia University

Outline

1. Model Output Statistics (MOS): bias correction, calibration.

2. Local calibration techniques. Examples.

3. Non-local calibration techniques. Examples.

4. Exercises: pattern-based calibration with PyCPT

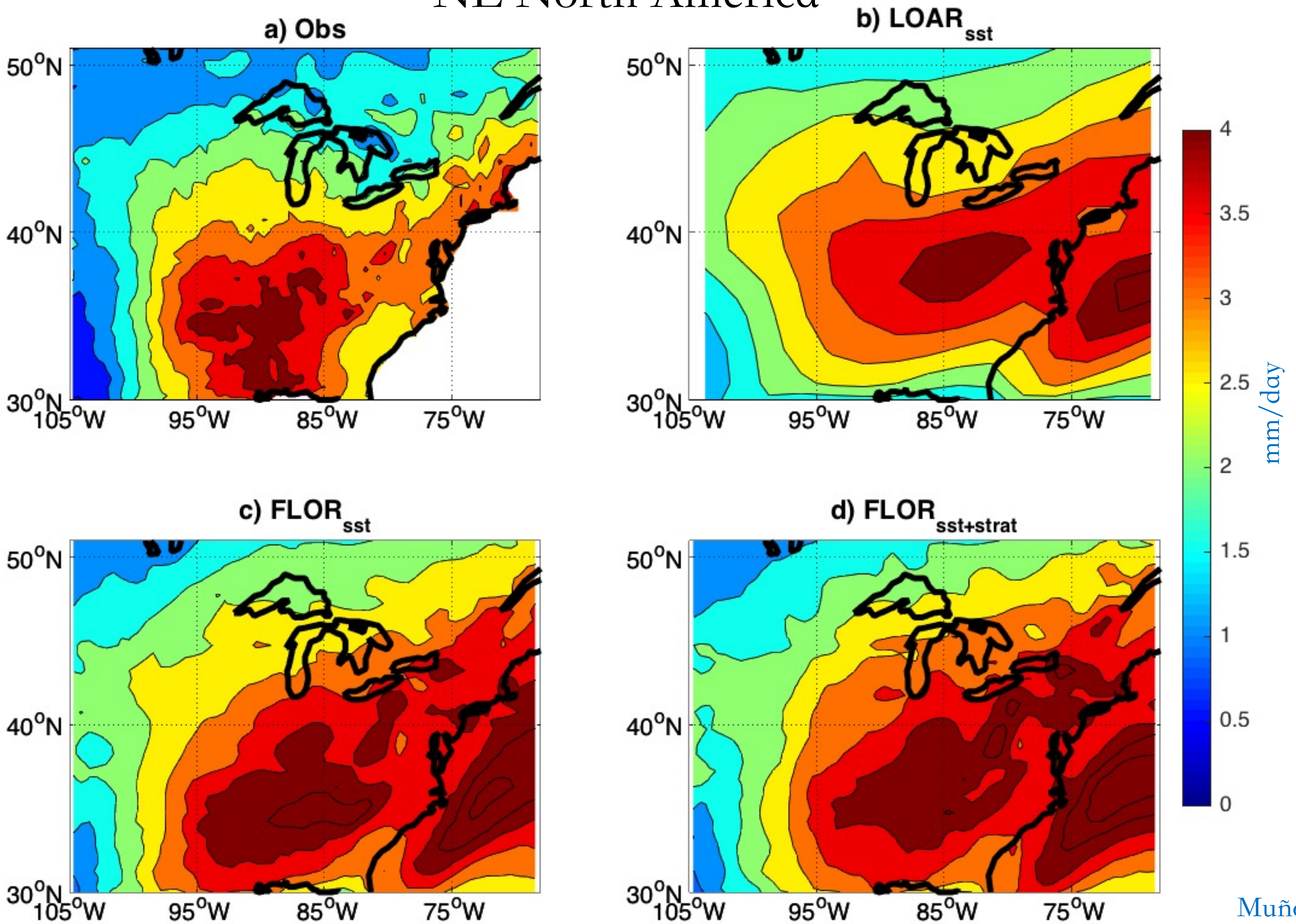
Outline

1. Model Output Statistics (MOS): bias correction, calibration.
2. Local calibration techniques. Examples.
3. Non-local calibration techniques. Examples.
4. Exercises: pattern-based calibration with PyCPT

Model vs Observations

Rainfall Climatology MAM – 1981-2012 – GFDL models

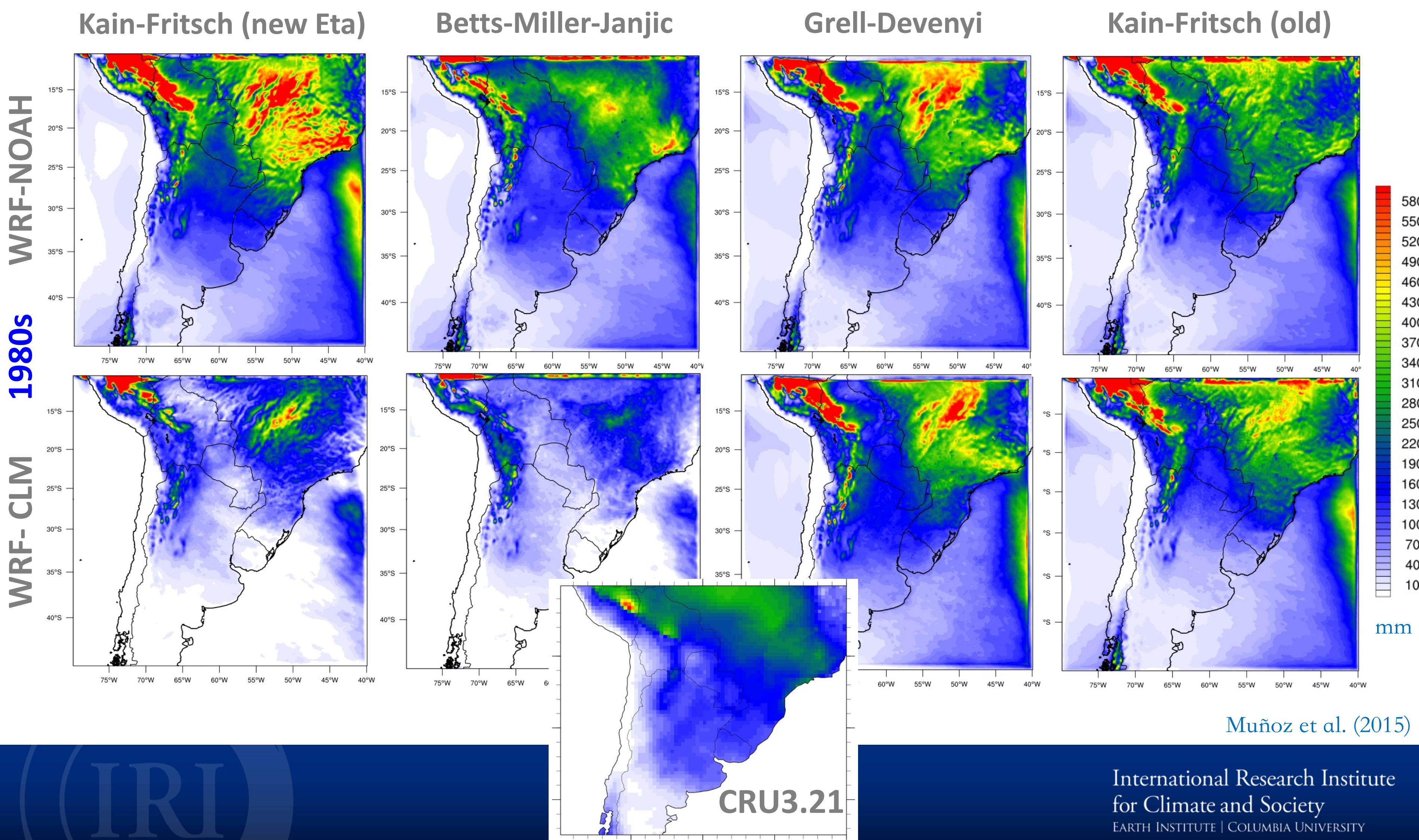
NE North America



Muñoz et al., 2017

Model vs Observations

Rainfall Climatology DJF – 1981-1989 – WRF model
SE South America



Muñoz et al. (2015)

Model Output Statistics

- Because of uncertainties in initial/boundary conditions, unknown or unresolved physical processes and the chaotic nature of the climate system, ***models are always subject to error***.

- Part of those errors are systematic, and can be corrected using **Model Output Statistics (MOS)**.

- Other errors are *not* correctible, and it is customary to provide an ensemble forecast to quantify uncertainties. This leads to probabilistic forecasts.

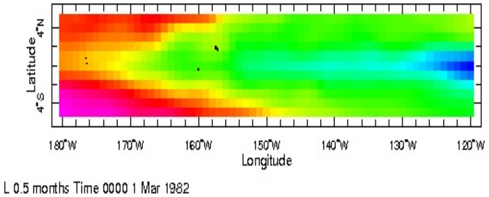
Model Output Statistics

- Generally speaking, MOS is any postprocessing we conduct on the raw model output (not only bias correction).
- In particular, it refers to statistical postprocessing to correct model errors.
- There are different types of biases, and different methods to correct them! Not a "unique MOS"…
- In this talk, to simplify language, any postprocessing method that corrects/calibrates model outputs are referred to as MOS.

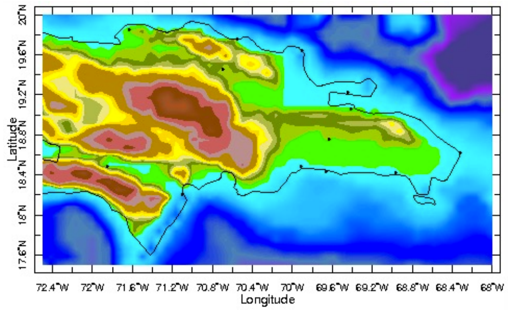
Calibrated Forecasts

Predictor (X) | Predictand (Y) | Calibrated Forecast

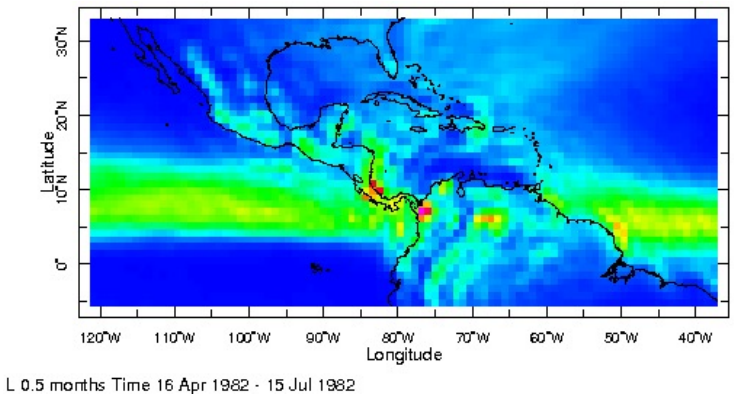
Empirical approach



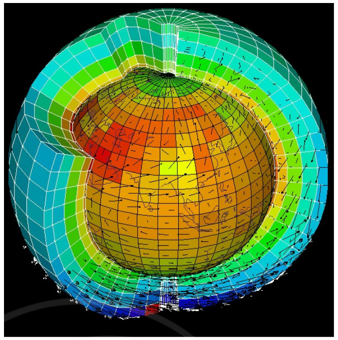
Obs



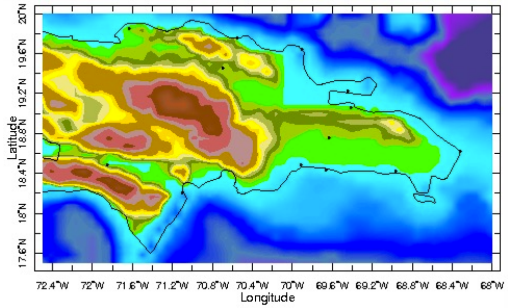
Obs



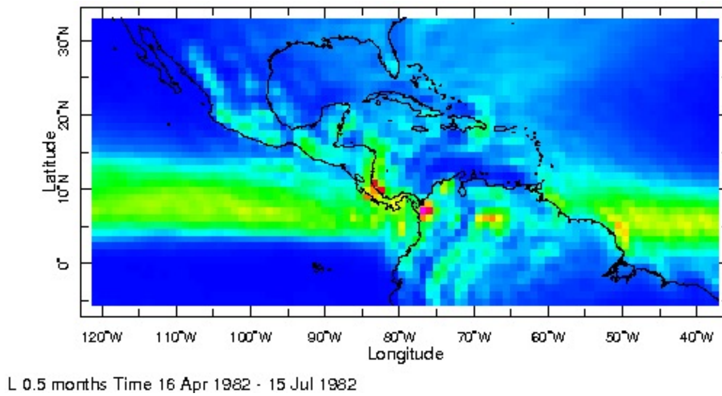
Hybrid approach



Dynamical Model



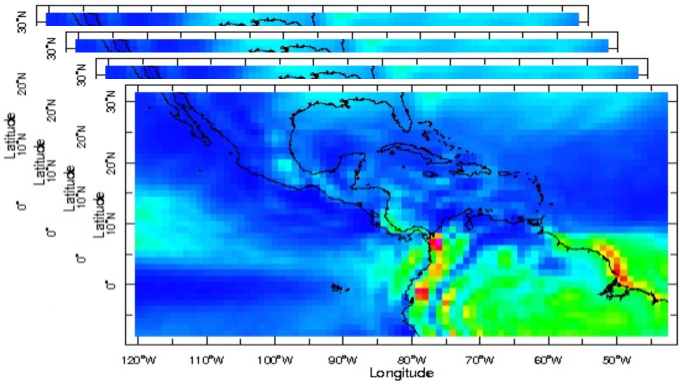
Obs



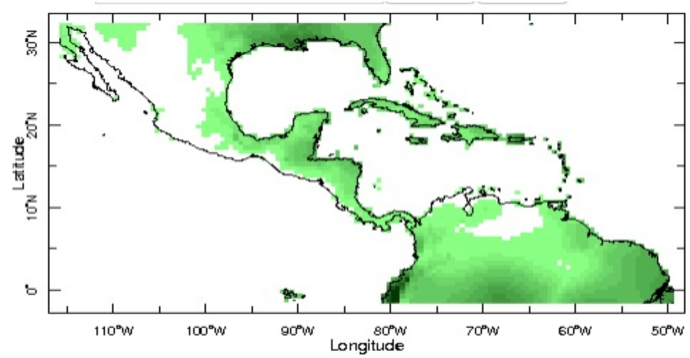
The NextGen Approach

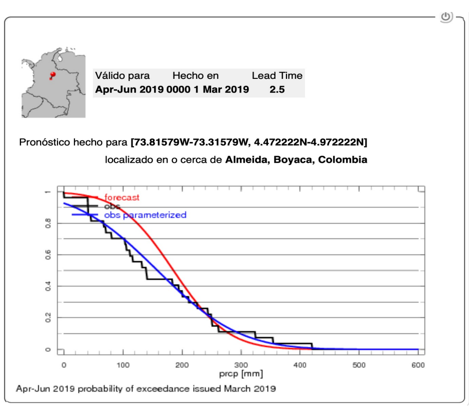
Raw Model Output (Predictors)

Multiple Obs and (C)GCMs [NMME, C3S, LC-LRFMME]

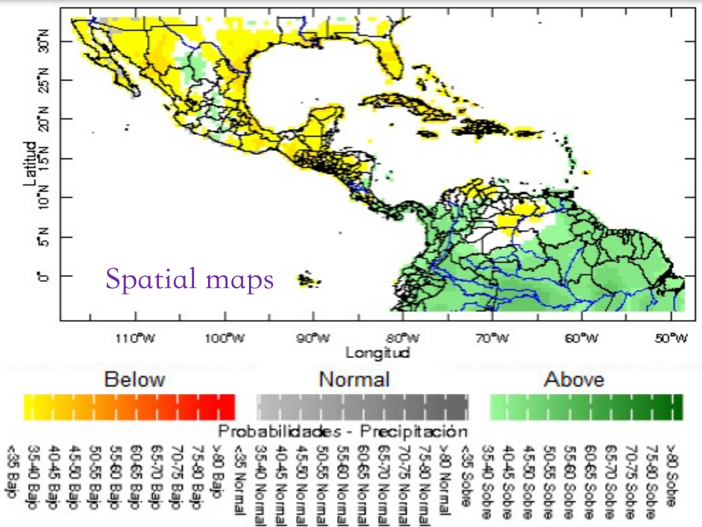


Gridded/Station Obs (Predictand)





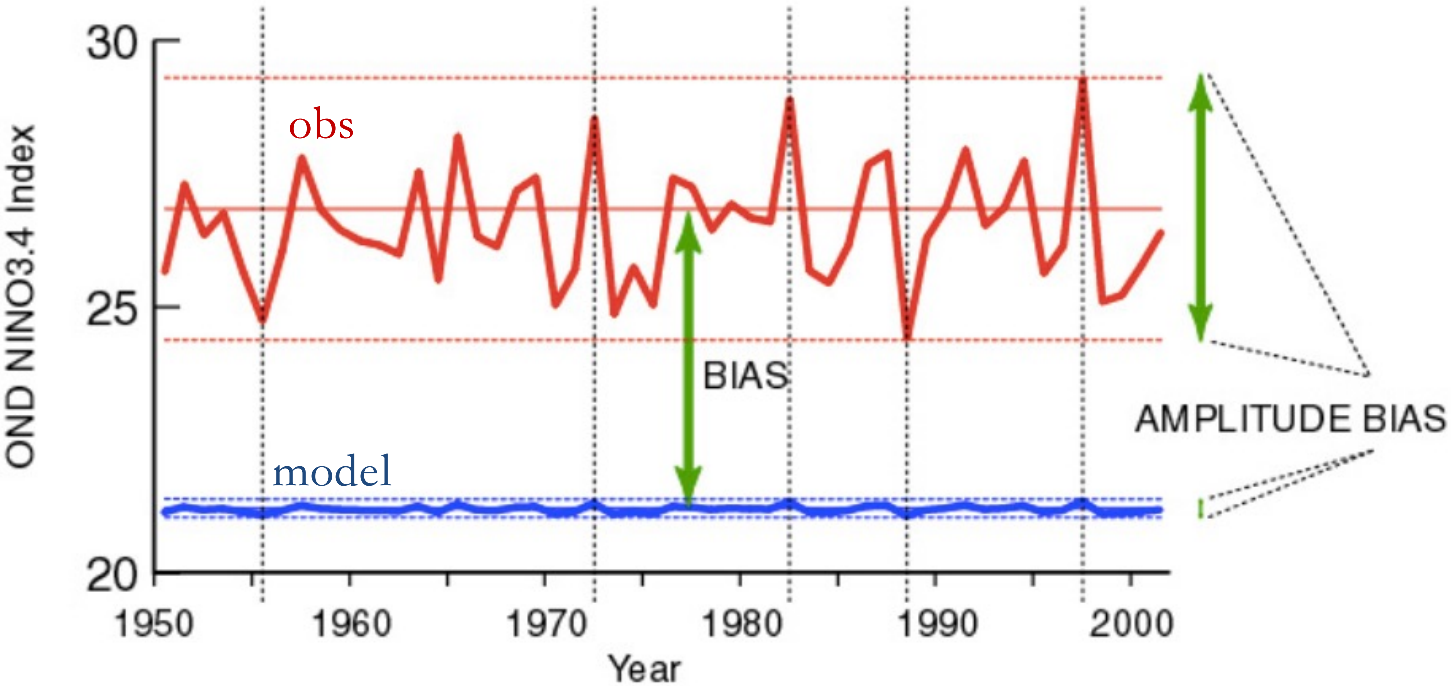
Forecast the entire PDF!



Pattern-based calibration (Model Output Statistics), each model independently, then ensemble in probability space

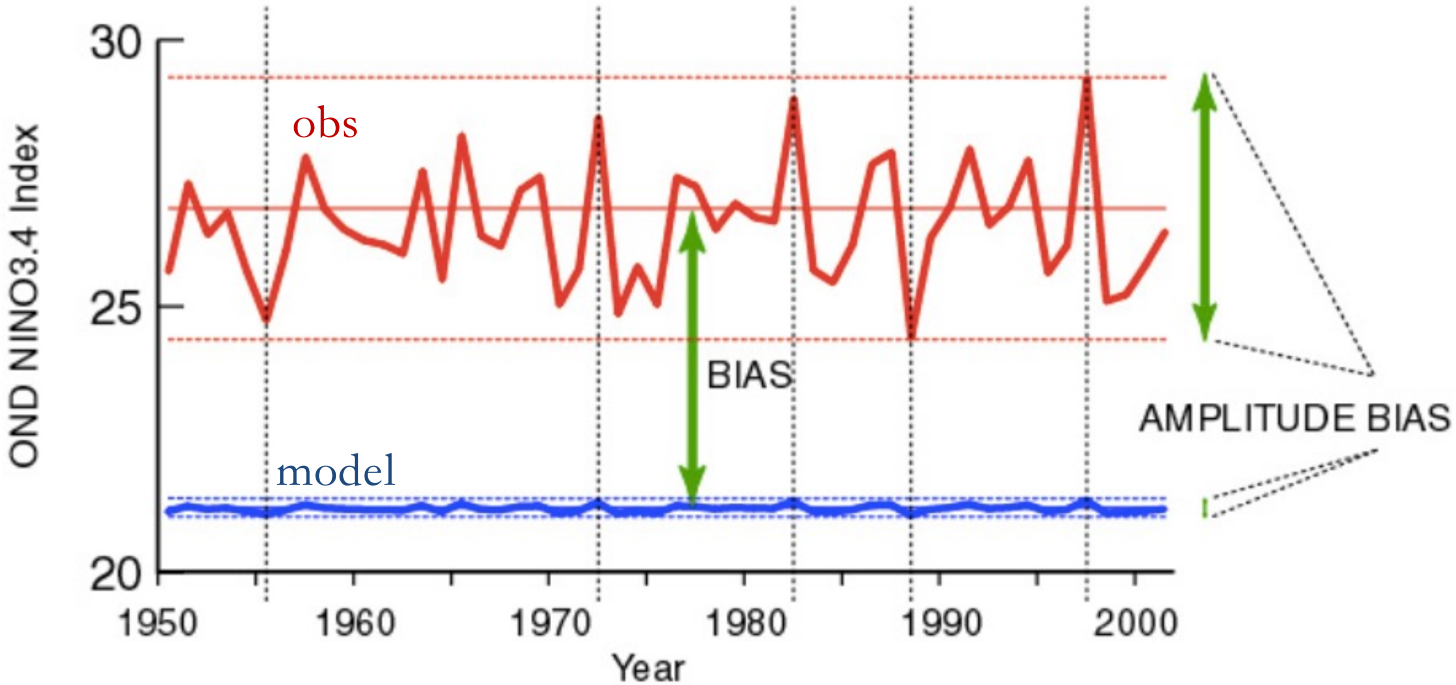
Model Output Statistics

- It is common to use Anomaly Correlation Coefficient to assess forecast skill, but it only measures *association*.
- There are **a lot of other forecast attributes** of interest! (remember Caio's class).

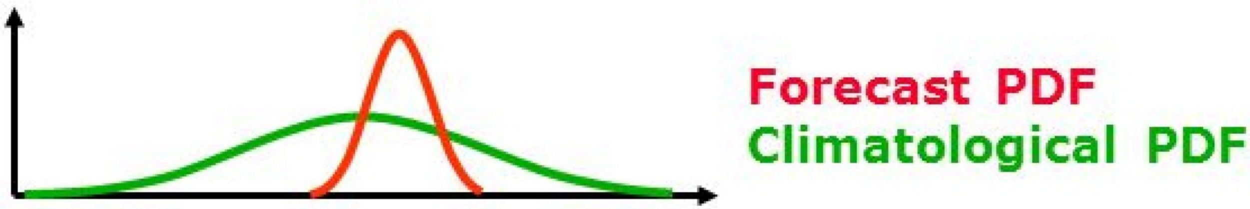


Courtesy of Simon J. Mason

Model Output Statistics

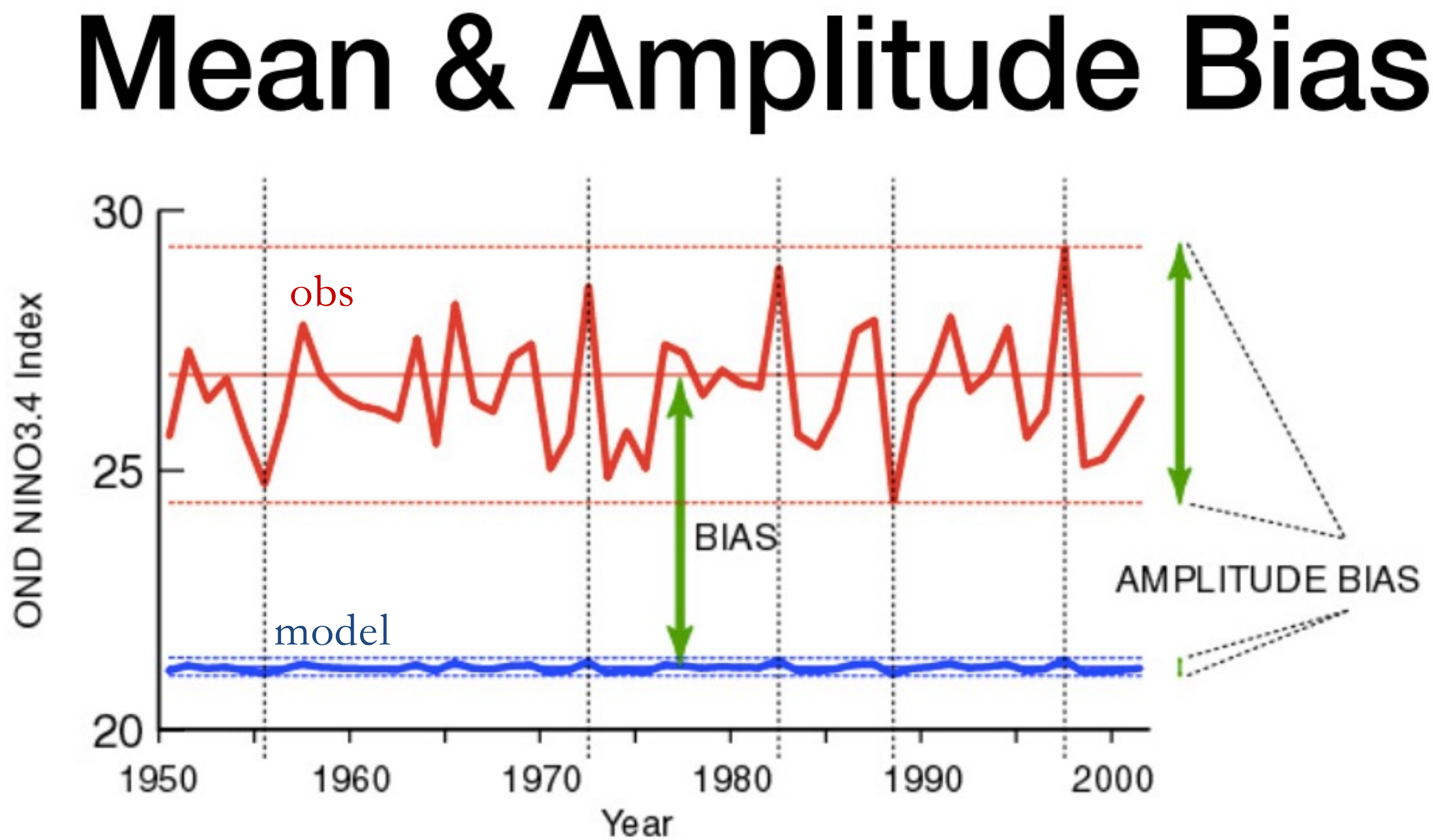


Courtesy of Simon J. Mason



Courtesy of Renate Hagedorn

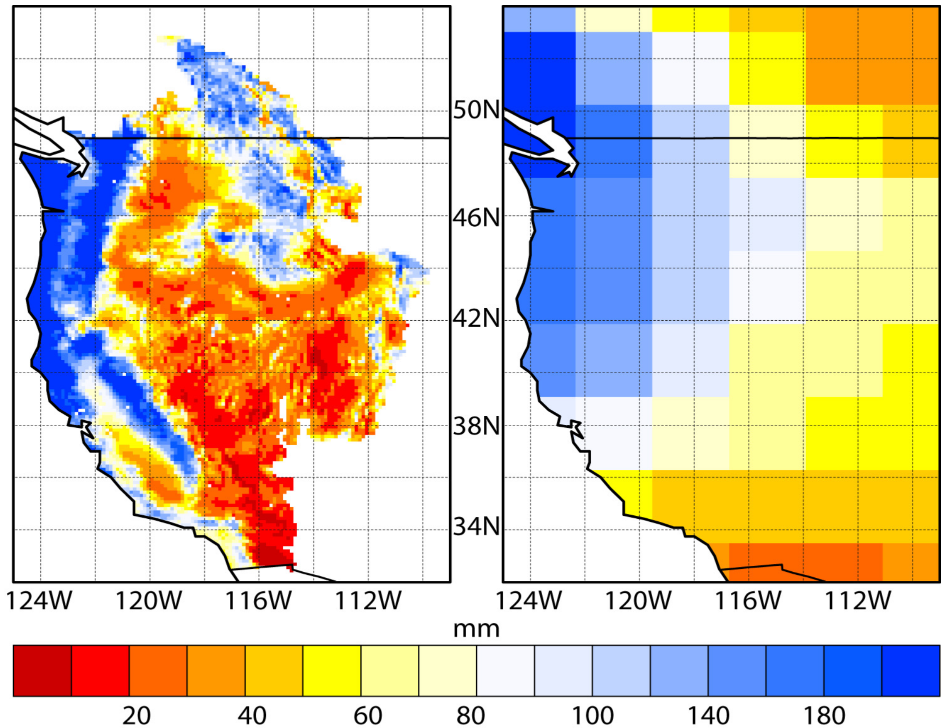
Which Biases to take care of?



Courtesy of S. Mason

Which Biases to take care of?

Local biases may have obvious reasons



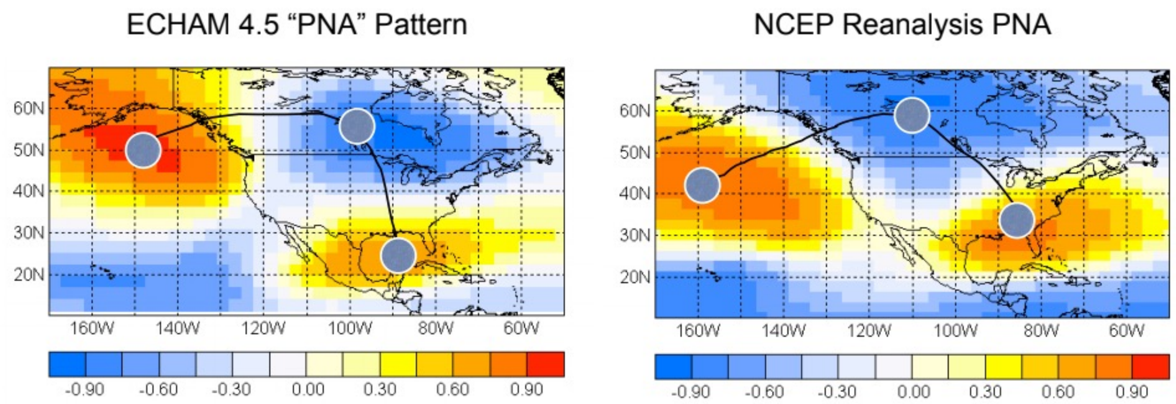
Annual mean observed and simulated precipitation over western North America.

Courtesy of Andrew W. Robertson

Climate can vary dramatically over short distances, especially in the context of precipitation and wind speeds.

Which Biases to take care of?

Conditional Bias
(errors in patterns of variability)



Important climate features may be displaced in GCMs relative to observations: *Systematic spatial biases*

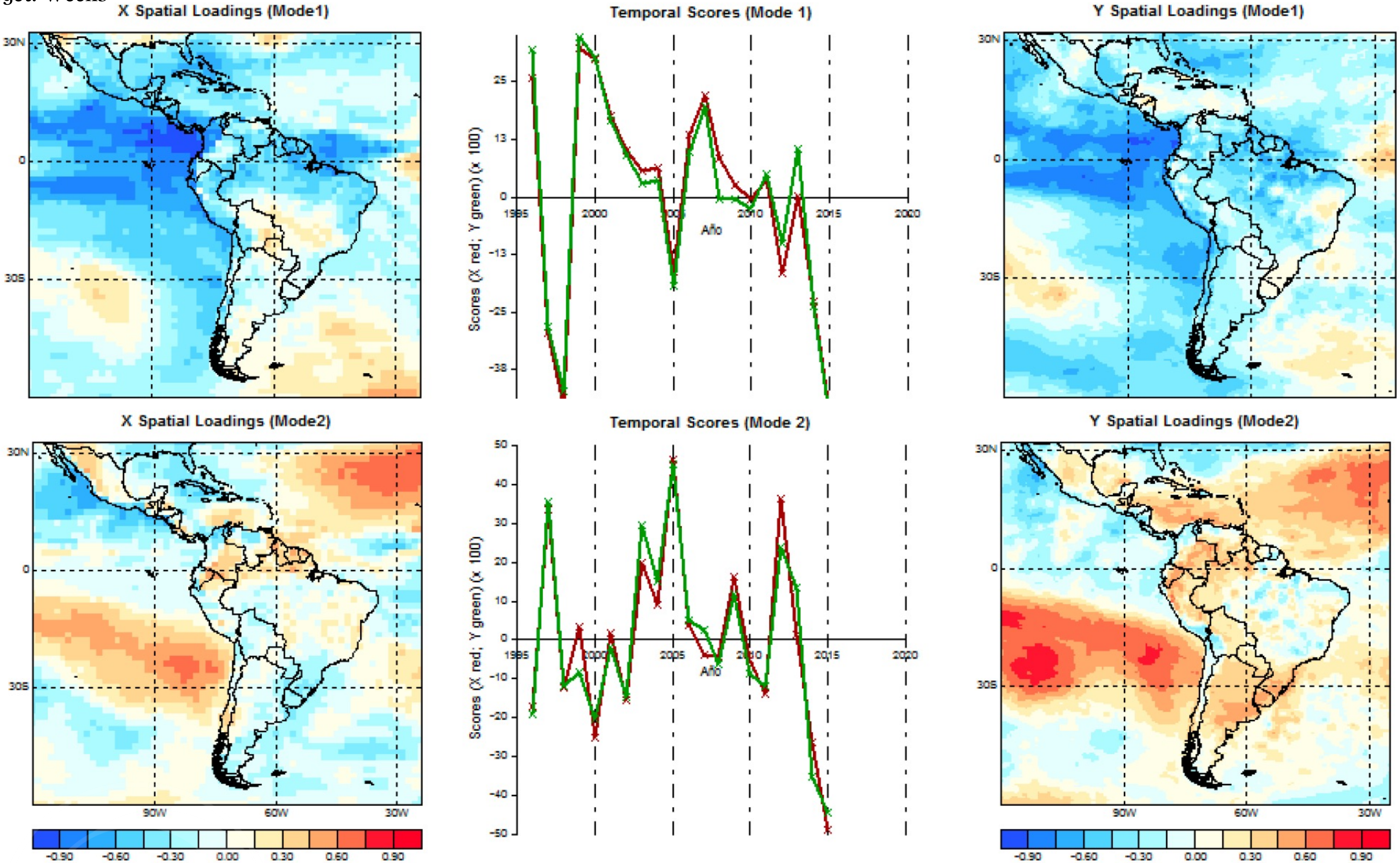
Courtesy B. Lyon (or was it Simon?)

Example:
T2M
Init: June
Target: Week3

Another example: T2M

Obs

Model



Negative MSSS implies that conditional biases can be LARGE.
Need to correct them!

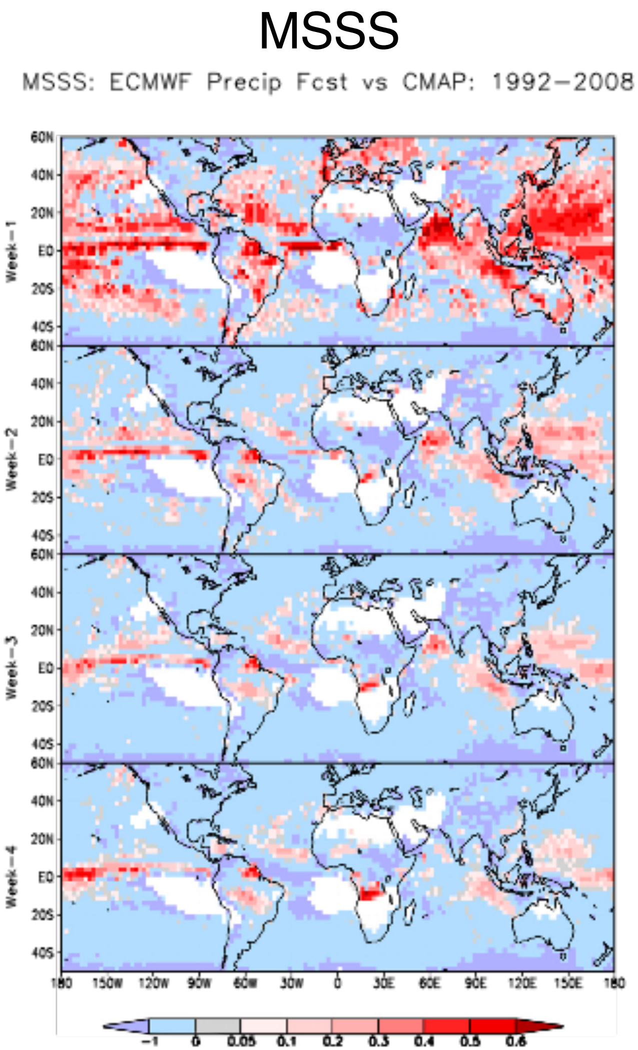
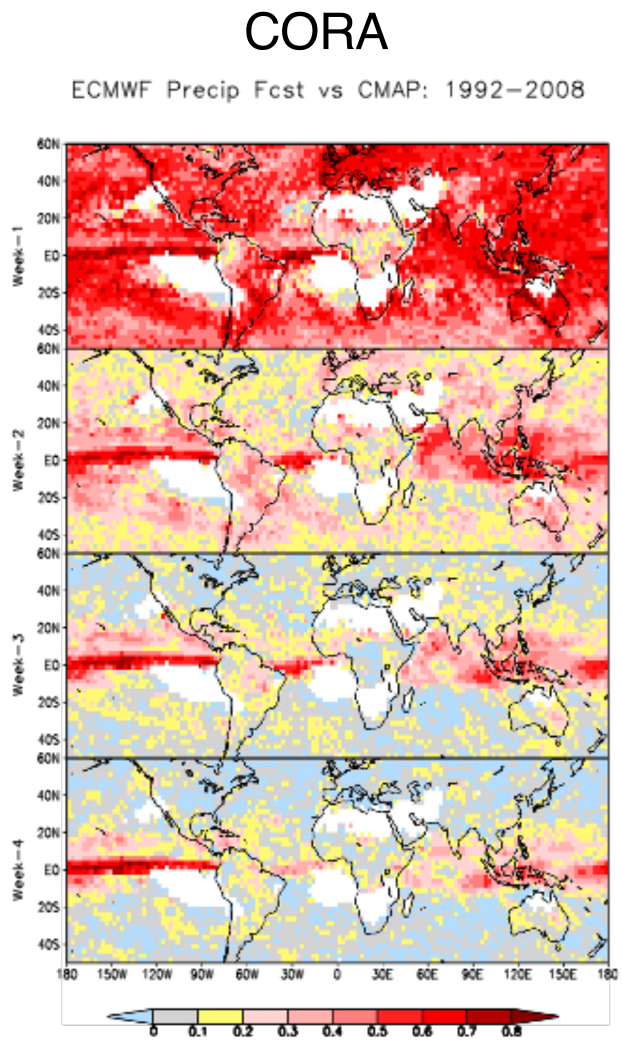


FIG. 15. Mean square skill score (MSSS) between ECMWF precipitation hindcast and CMAP rainfall data over weeks 1–4.



Li and Robertson, 2015

The MSSS can be decomposed into the square of the correlation coefficient, together with the squares of the conditional and mean prediction biases (Murphy 1988). If the mean biases are subtracted at the outset, the MSSS consists of only the $(\mathrm{CORA})^2$ and the squared conditional bias.

$$\mathrm{MSSS} = (\mathrm{CORA})^2 - (\text{conditional bias})^2$$

Outline

1. Model Output Statistics (MOS): bias correction, calibration.

2. Local calibration techniques. Examples.

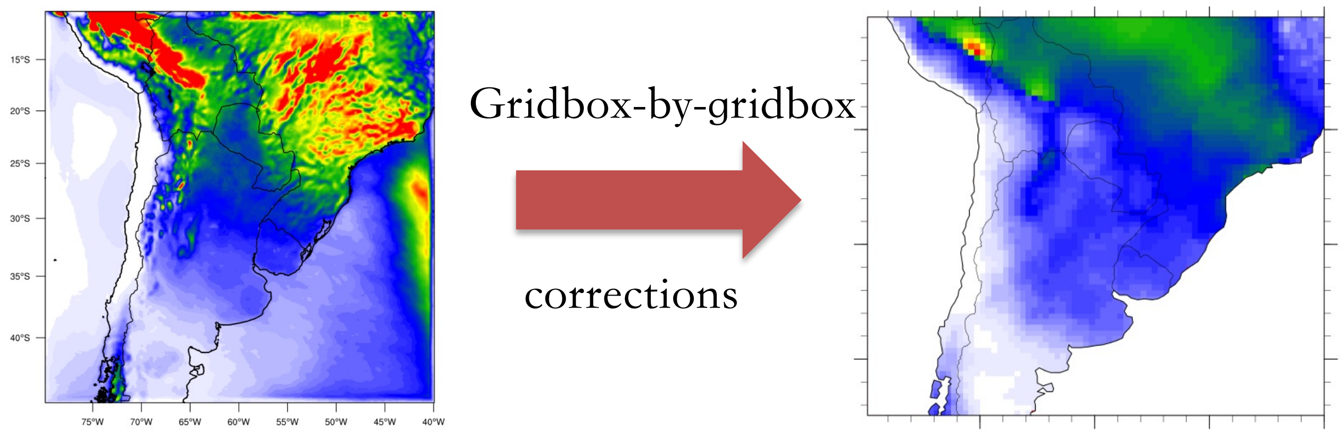
3. Non-local calibration techniques. Examples.

4. Exercises: pattern-based calibration with PyCPT

Local Calibration Techniques

- Post-processing gridbox by gridbox so the desired statistical characteristics in the model "match" the observed ones.

- Sometimes considered a "brute force" approach (in the sense that it is usually not informed by physics, just statistics).



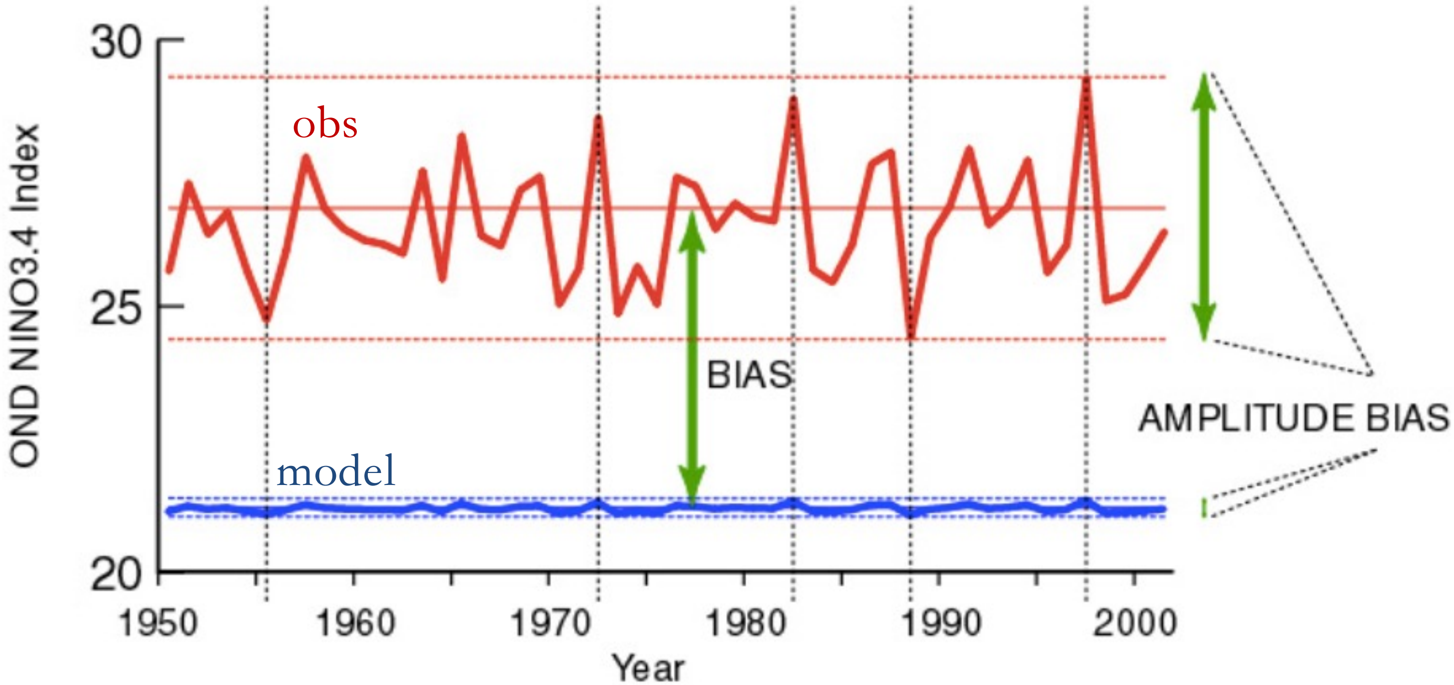
Model

Obs

Local Calibration Techniques

- Typically addressing mean and amplitude biases, but there are methods that go well beyond those corrections.

- Matching spatial and temporal resolution of datasets can be important.



Local Calibration Techniques

Let's see some methods (from Manzanas et al., 2019)

All the methods described in this section have been applied gridbox by gridbox considering seasonal interannual series. We use the following notation: $y_{m,t}$ and $y'_{m,t}$ denote the original and calibrated values for the ensemble member m at time (season/year) t, $\hat{y}$ is the average of the ensemble mean ($\bar{y}_t$) on all times t, $\hat{o}$ is the average of the observations on all times t, σ_f is the standard deviation of the complete ensemble (pooling all member interannual time-series) and σ_o is the standard deviation of the observed interannual time-series. Finally, ρ is the interannual correlation between the ensemble mean and the observational reference.

Local Calibration Techniques

Mean and Variance Adjustment (MVA) – e.g., Leung et al. 1999

Mean correction

$$y'_{m,t} = (y_{m,t} - \hat{y}) \qquad + \hat{o}$$

Mean and amplitude correction

$$y'_{m,t} = (y_{m,t} - \hat{y})\frac{\sigma_o}{\sigma_f} + \hat{o}$$

Local Calibration Techniques

Empirical Quantile Mapping (EQM) – e.g., Manzanas et al., 2018

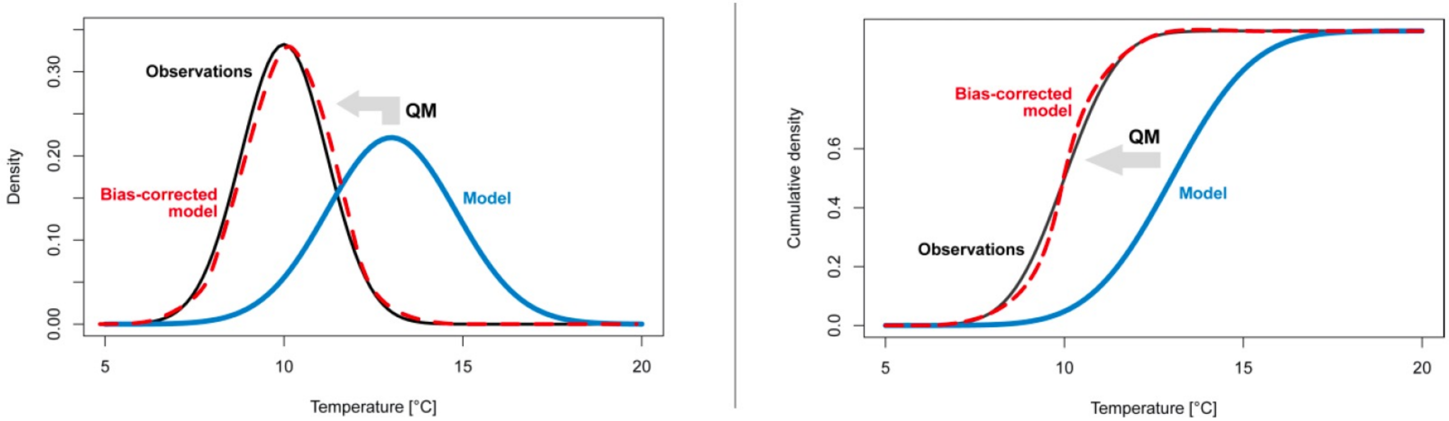


Figure 3: The nature of QM: A biased simulated distribution (blue) is corrected towards an observed distribution (black). In the example shown the raw simulated distribution is subject to both a bias of the mean and a bias in variance. The resulting bias-corrected distribution (dashed red) approximates the observed one but is typically not identical to it (e.g. due to the sampling uncertainty during the calibration of the correction function or details of the specific QM implementation). Left panel: Example based on the probability density function (PDF). Right panel: example based on the cumulative distribution function (CDF).

Technical Report MeteoSwiss No. 270

Local Calibration Techniques

Climate conserving realibration (CCR) – e.g., Doblas-Reyes et al., 2005

$$y'_{m,t} = \rho \frac{\sigma_o}{std(\bar{y}_t)} \bar{y}_t + \sqrt{1-\rho^2} \frac{\sigma_o}{\sigma_f} (y_{m,t} - \bar{y}_t) + \hat{o}$$

- Corrects forecasts so they have the observed interannual variance
- Preserves inter-annual correlation

Local Calibration Techniques

Ratio of predictable components (RPC) – e.g., Eade et al., 2014

$$y'_{m,t} = \rho \frac{\sigma_o}{std(\bar{y}_t)}(\bar{y}_t - \hat{y}) + \sqrt{1-\rho^2}\frac{\sigma_o}{\sqrt{var(y_{m,t} - \bar{y}_t)}}(y_{m,t} - \bar{y}_t) + \hat{o}$$

- Uses ensemble to reduce noise
- Adjust forecast variance to ratio of predictable components in the model matches the observed one

Local Calibration Techniques

Correction/downscaling of GCM forecasts

- If we have a set of GCM hindcasts (x) and verifying observations (y), then we can build a regression model to relate them.

$$\hat{y} = \beta_0 + \beta_1 x_1$$

- The GCM forecasts becomes the “predictor” x in the regression model, and the observations becomes the “predictand” y

- Regression models trained on GCM hindcasts vs historical data are called “MOS Correction” (for Model Output Statistics)

- Once the regression coefficients have been estimated from hindcasts (e.g., past 20 years), the model can be used to correct a new forecast x for t=2021.

Local Calibration Techniques

Linear Regression (LR) – e.g., Manzanas et al., 2019

$$o_t = \alpha + \beta \bar{y}_t + \epsilon$$

$$y'_{m,t} = \alpha + \beta \bar{y}_t + \gamma_t (y_{m,t} - \bar{y}_t)$$

$$\gamma_t = std(\epsilon_{fit}) \sqrt{1 + 1/n + \frac{(y_t - \bar{y}_t)^2}{(n-1)var(\epsilon_{obs})}}$$

- Simple linear regression between the ensemble mean and observations
- Adjust forecast variance by rescaling the standard deviation of the predictive distribution from the linear fit

Local Calibration Techniques

Non-homogeneous Gaussian Regression (NGR) – e.g., Gneiting et al., 2005

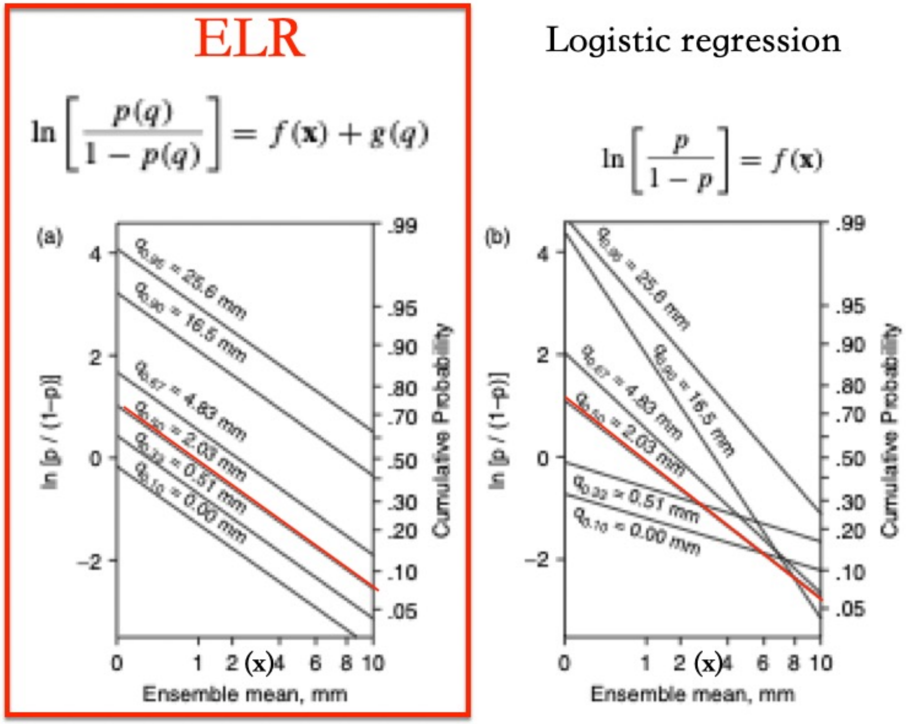
$$y'_{m,t} = \alpha + \beta(\bar{y}_t - \hat{y}) + \sqrt{\gamma^2 + \delta^2 var(y_t)}(y_{m,t} - \bar{y}_t)$$

- Ensemble mean signal and spread (+constant) are used as predictors for the calibrated forecast mean and variance, respectively.
- Parameters are optimized by minimizing the ensemble CRPS

Local Calibration Techniques

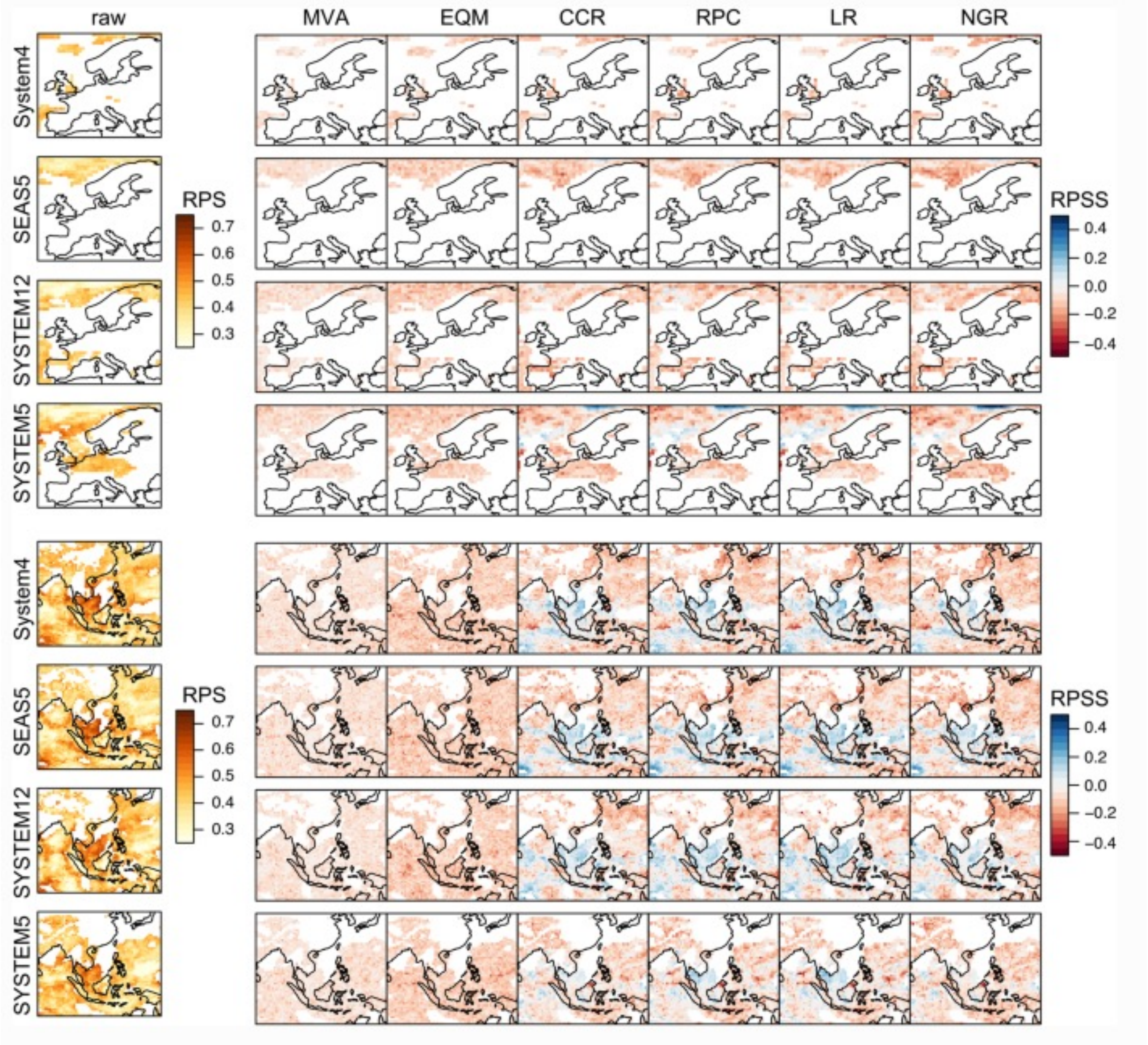
Extended Logistic Regression (ELR) – e.g., Vigaud, Robertson and Tippet, 2017

GFS Day 6 – 10
Precip Forecast for
Minneapolis
28 Nov – 2 Dec
2001
Wilks (2009)



Applied at each grid point, using forecast ensemble mean

Local Calibration Techniques



Manzanas et al 2019

Outline

1. Model Output Statistics (MOS): bias correction, calibration.

2. Local calibration techniques. Examples.

3. Non-local calibration techniques. Examples.

4. Exercises: pattern-based calibration with PyCPT

Non-Local Calibration Techniques

Remember our regression equation?

- If we have a set of GCM hindcasts (x) and verifying observations (y), then we can build a regression model to relate them.

$$\hat{y} = \beta_0 + \beta_1 x_1$$

- The GCM forecasts becomes the "predictor" x in the regression model, and the observations becomes the "predictand" y

- Regression models trained on GCM hindcasts vs historical data are called "MOS Correction" (for Model Output Statistics)

- Once the regression coefficients have been estimated from hindcasts (e.g., past 20 years), the model can be used to correct a new forecast x for t=2021.

Non-Local Calibration Techniques

Varieties of linear regression

Local

- simple regression: a **univariate** predictor and a **univariate** predictand:
  $y = ax + b$

- multiple regression: **two or more** predictors, and a **single** predictand
  $y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n$ (case of n predictors)

  -- e.g., **Principal Components Regression (PCR)**

- multivariate (pattern) regression: **two or more** predictors, **two or more** predictands
  $\mathbf{y} = \mathbf{Ax} + \mathbf{b}$ (matrix A)

  -- e.g., **Canonical Correlation Analysis (CCA)**

Non-local

Non-Local Calibration Techniques

Canonical Correlation Analysis

- To predict the observed precip anomaly field **y** from a GCM forecast anomaly field **x**, we assume the linear relationship $\mathbf{y} = \mathbf{A}\mathbf{x}$
- We minimize the regression error squared $<(\mathbf{y} - \mathbf{A}\mathbf{x})^T(\mathbf{y} - \mathbf{A}\mathbf{x})>$
  by estimating the matrix **A** from hindcasts: $\mathbf{A} = (\mathbf{y}\mathbf{x}^T)(\mathbf{x}\mathbf{x}^T)^{-1}$
- This cannot be done when hindcast series is smaller than the spatial dimension of **x** and **y**
- The dimensions of **x** and **y** must be reduced!

Non-Local Calibration Techniques

Canonical Correlation Analysis

- Use Principal Component Analysis (PCA) to reduce the spatial dimension of the regression
- Truncate to typically < 10 PCs in **x** and **y**
- Data compression! ==> So # of predictors < # of training samples, which makes the multiple regression problem well-posed
- Big bonus: The PC time series are uncorrelated and maximize the variance. This solves the other problem with multiple linear regression when the predictors are correlated.

Non-Local Calibration Techniques

Confused? It's always the same idea…

Local

- simple regression: a **univariate** predictor and a **univariate** predictand:
  y = ax + b
- multipl… a **single** predict…
  y = a_0 … …ctors)

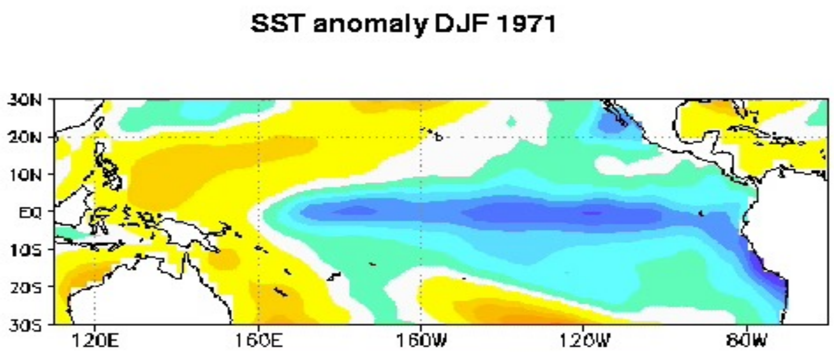
$$\hat{y} = \beta_0 + \beta_1 x_1$$

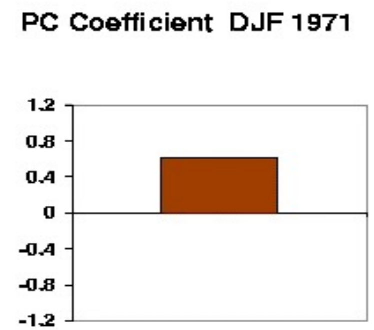
  -- e.g., **Principal Components Regression (PCR)**
- multivariate (pattern) regression: **two or more** predictors, **two or more** predictands
  y = Ax + b (matrix A)

Non-local

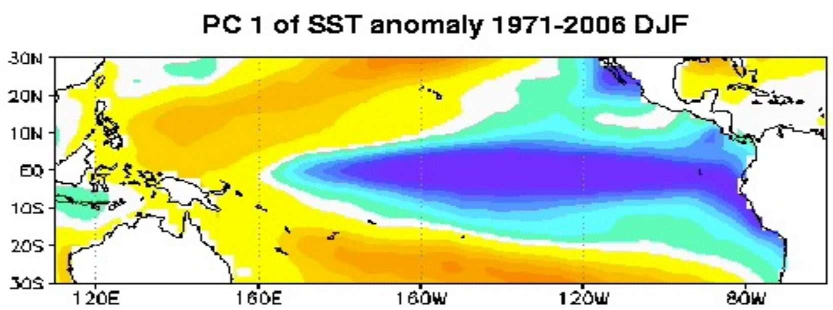
  -- e.g., **Canonical Correlation Analysis (CCA)**

Principal Component Timeseries as predictors



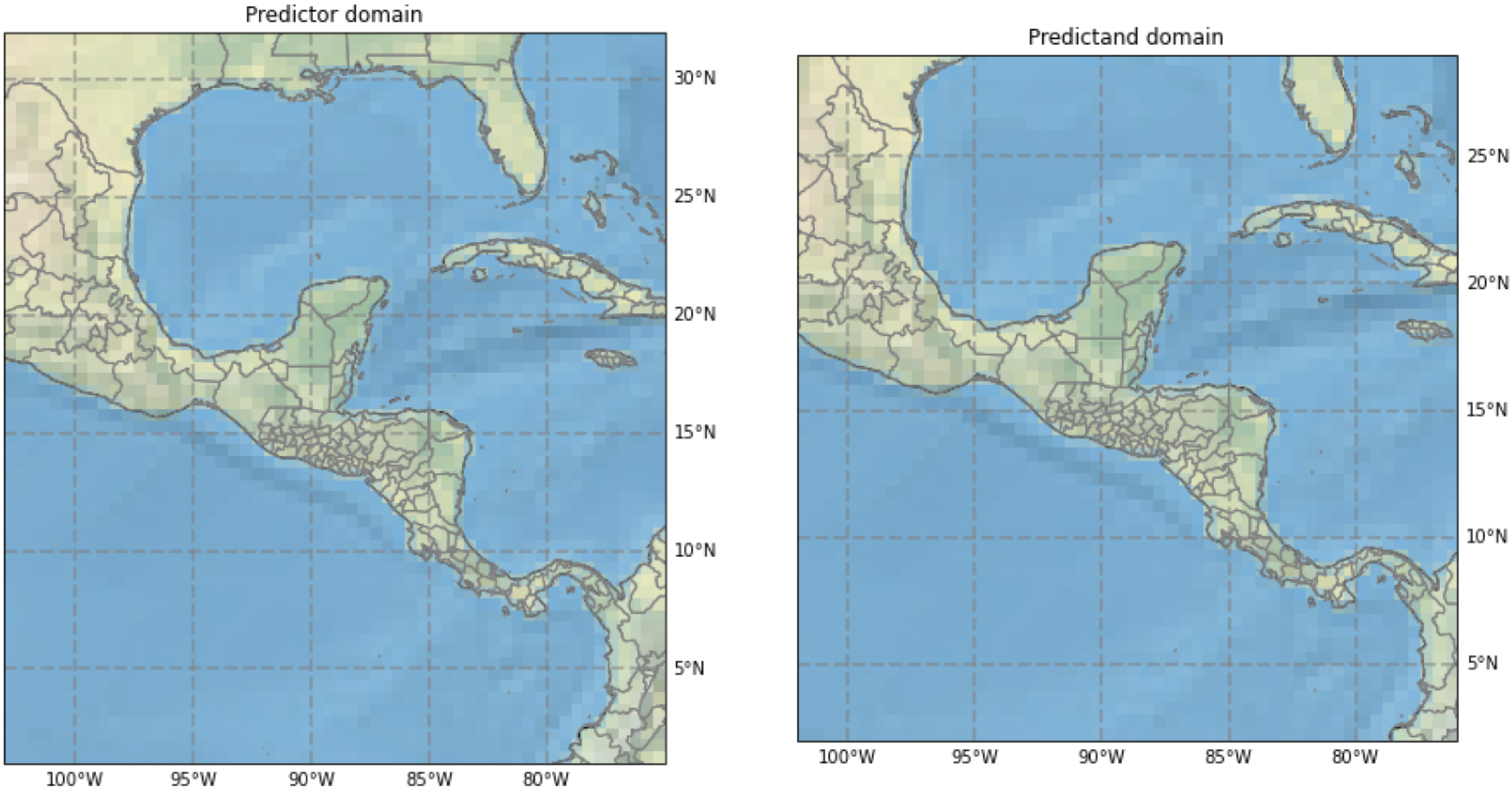


EOF 1



Courtesy of Ousmane Nyade

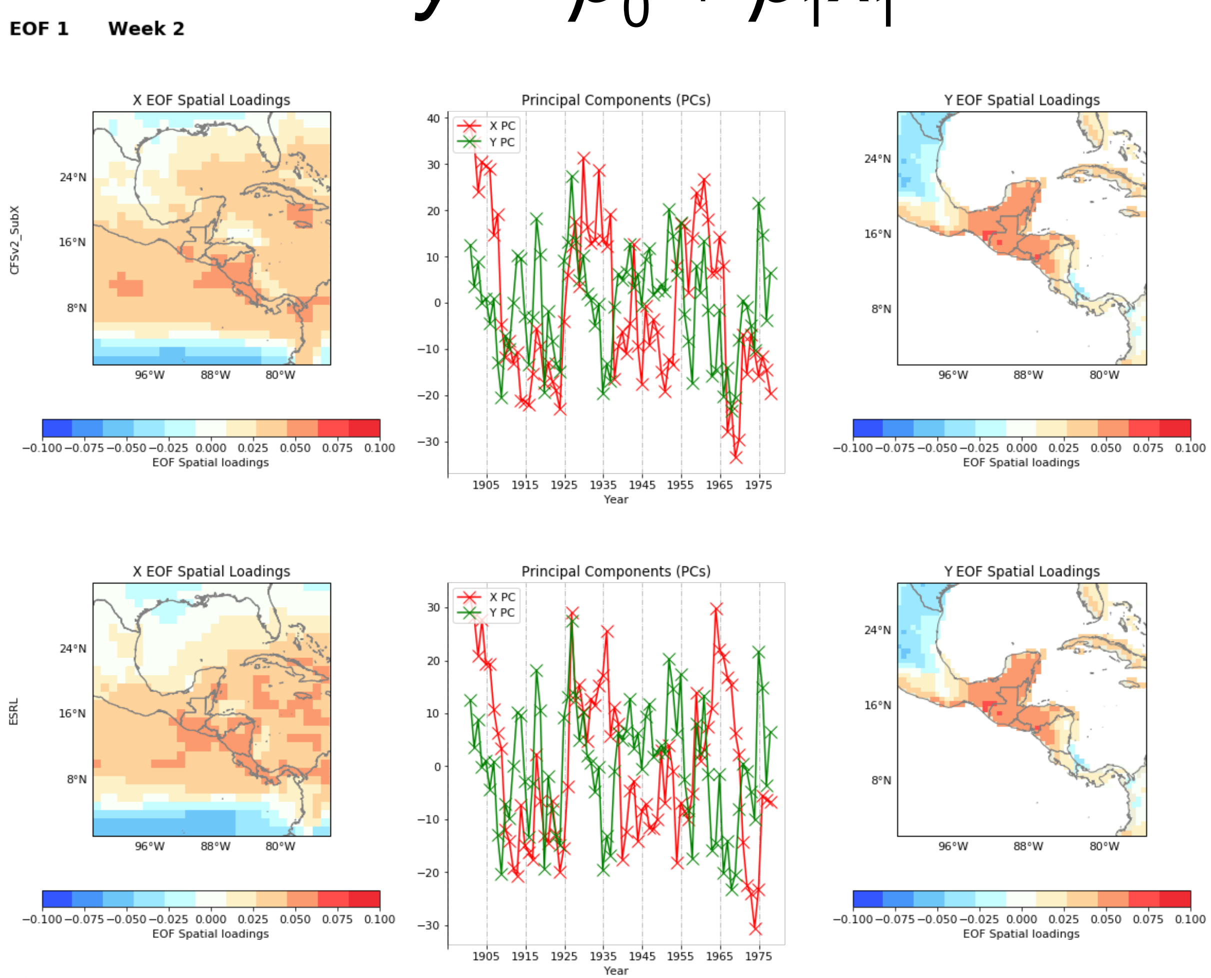
Non-Local Calibration Examples



Non-Local Calibration: PCR

$$\hat{y} = \beta_0 + \beta_1 x_1$$

EOF1

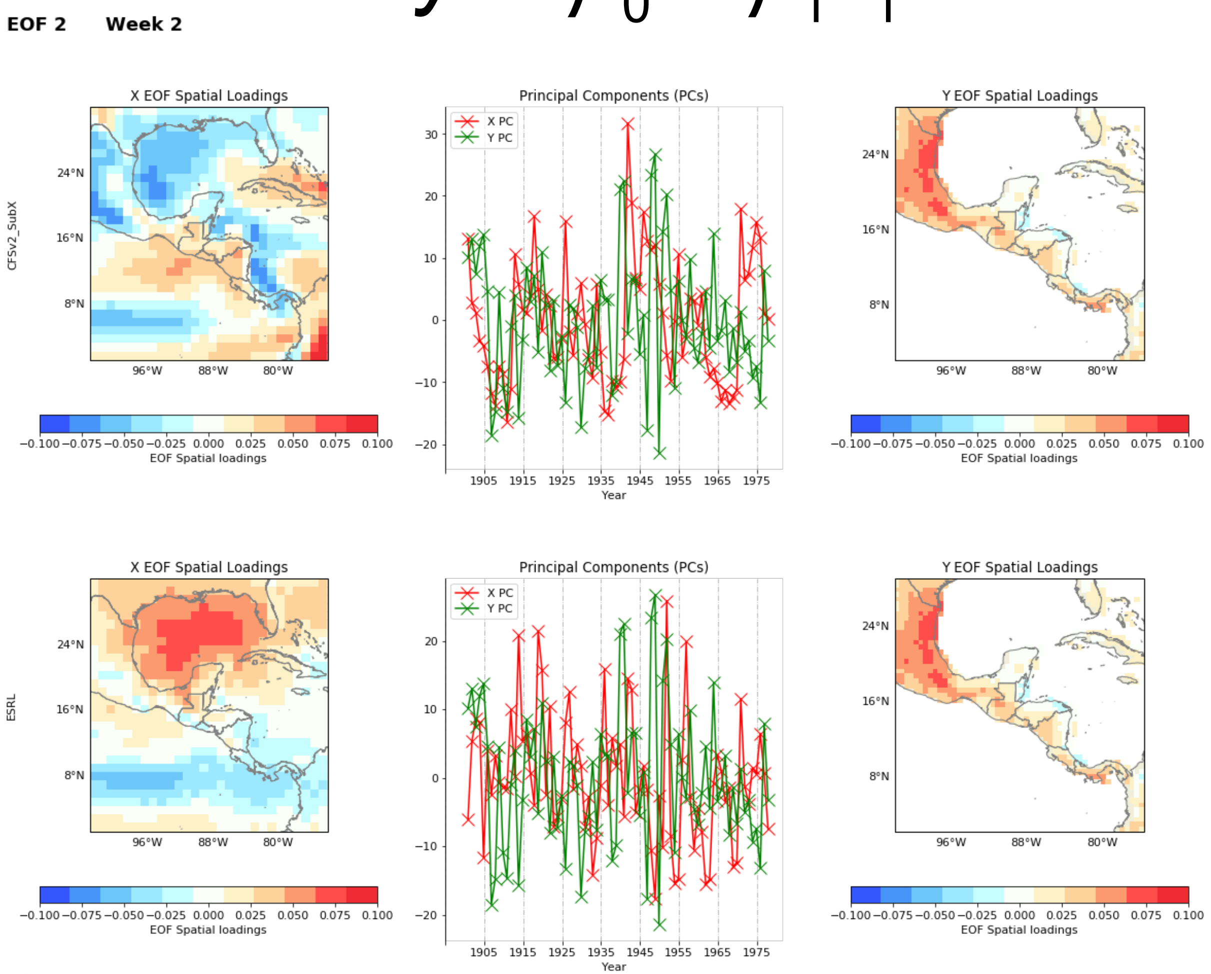


Each location is predicted as a linear combination of the model EOFs

Non-Local Calibration: PCR

$$\hat{y} = \beta_0 + \beta_1 x_1$$

EOF2



Each location is predicted as a linear combination of the model EOFs

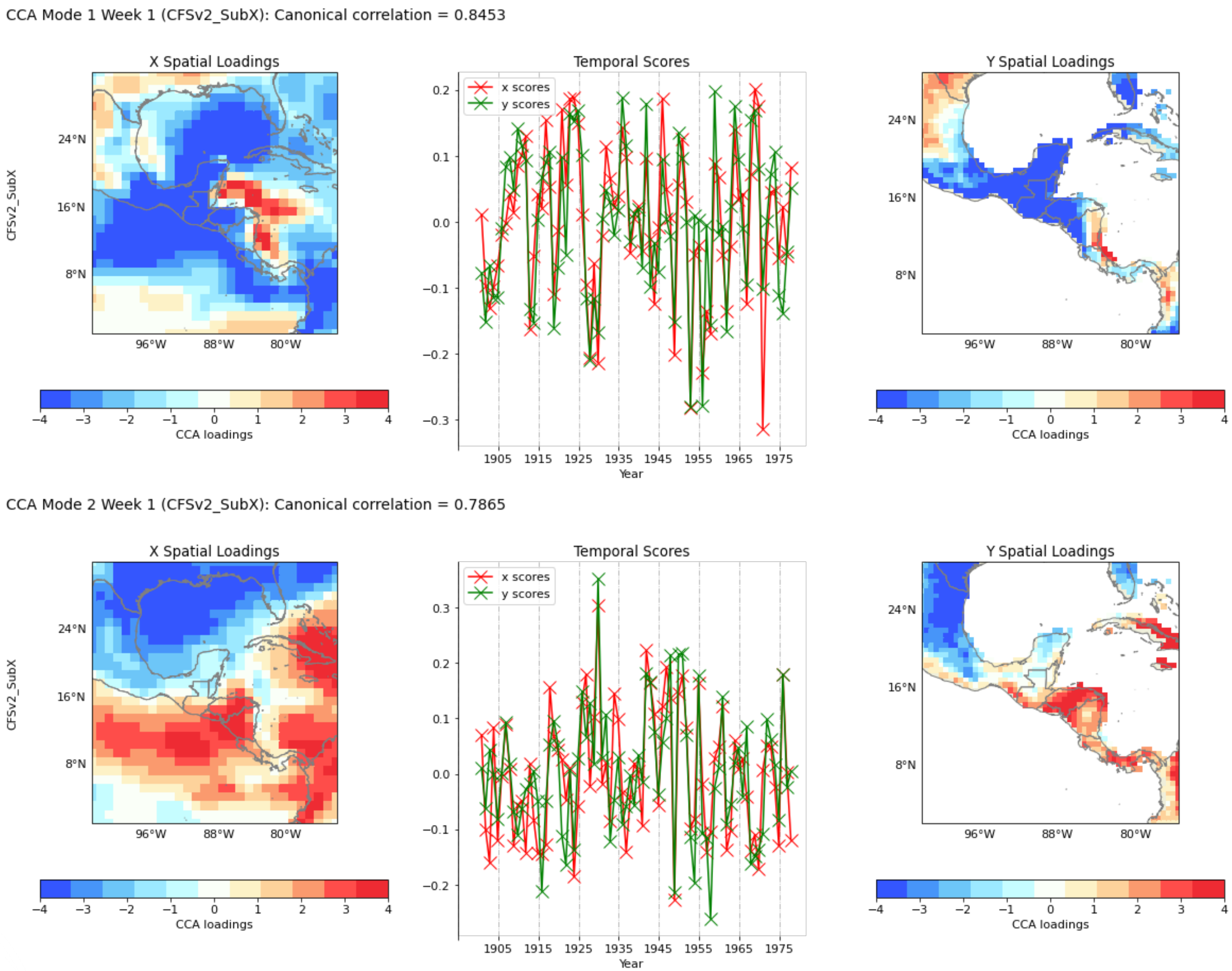
Non-Local Calibration: CCA

Linear combination of observed EOFs ⟺ Linear combination of model EOFs

The CCA modes are linear combinations of the EOFs of x and y such that their time series are maximally correlated

"Pattern regression" Corrects biases in patterns

How many PCs and CCA modes to retain?

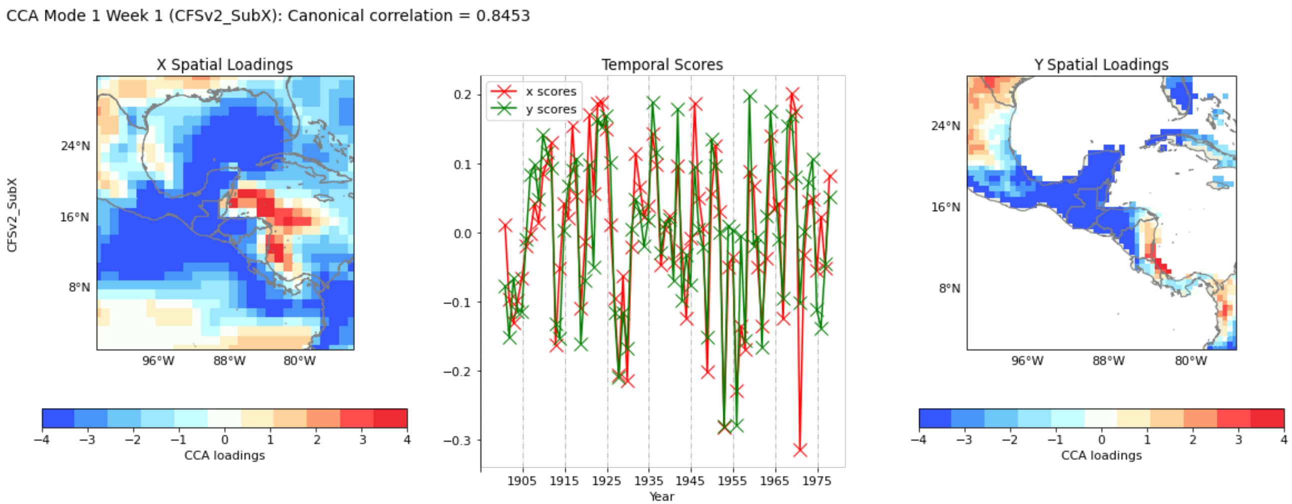


CCA modes
Same models

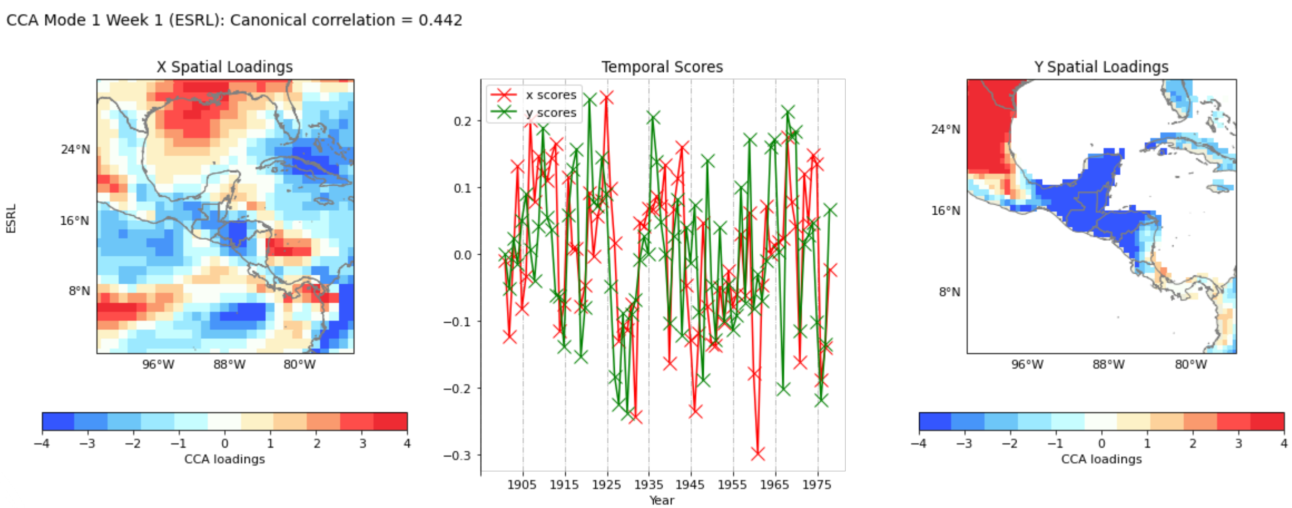
Non-Local Calibration: CCA

Linear combination of observed EOFs ⟺ Linear combination of model EOFs

The CCA modes are linear combinations of the EOFs of x and y such that their time series are maximally correlated



“Pattern regression”
Corrects biases in patterns



How many PCs and CCA modes to retain?

CCA mode 1
2 models

Comparing MOS methods

NoMOS Local → Non-local

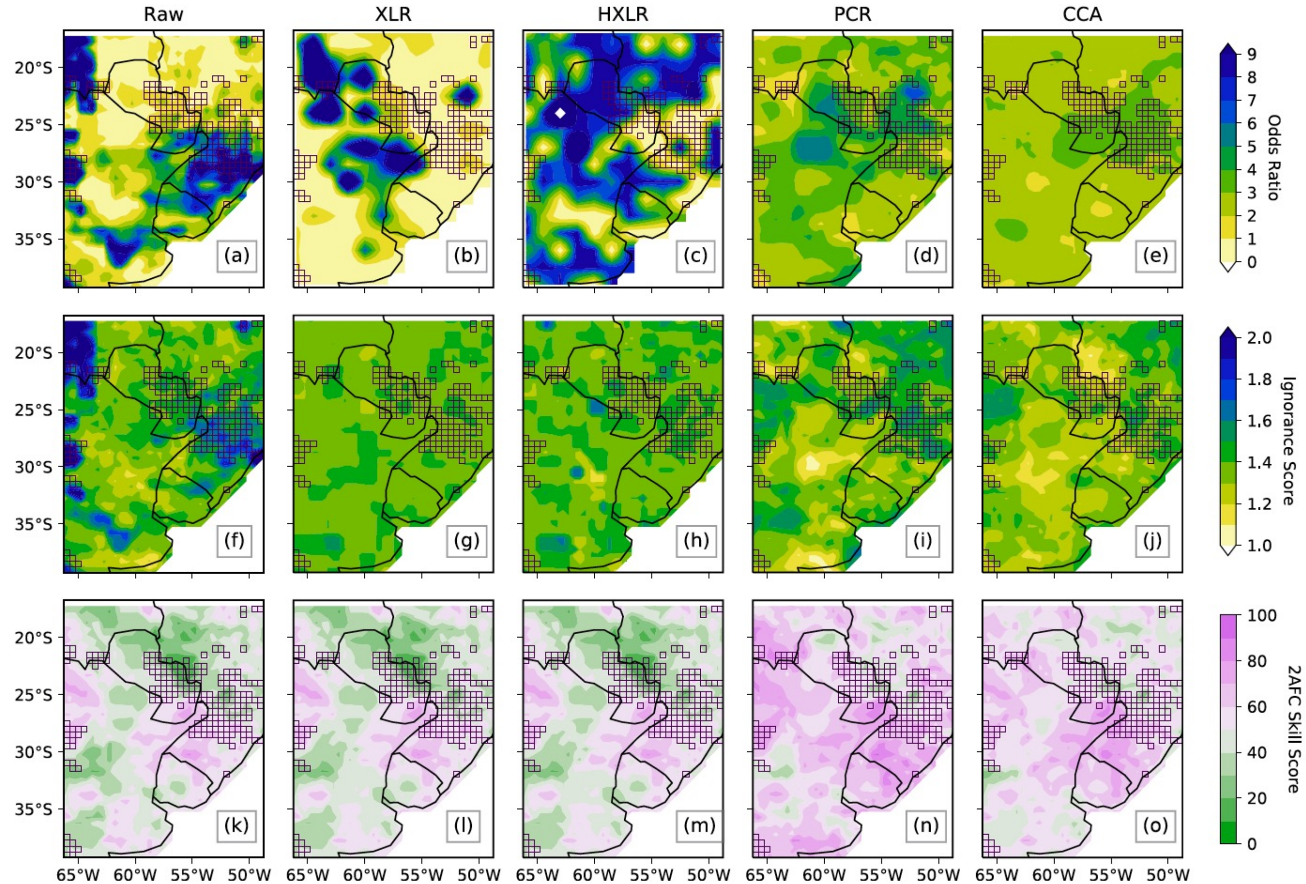


FIG. 10. Raw and MOS-adjusted S2S model forecasts and skill scores for the methods indicated in Table 1. (a)–(e) The heavy rainfall forecast for 1–7 Dec 2015 as odds_r defined in Eq. (3) over the target domain. A value greater than 1 indicates that the model forecast greater-than-average odds of rainfall exceeding the 90th percentile. (f)–(j) The IGN defined in Eq. (4), with zero indicating a perfect forecast. (k)–(o) The 2AFC skill score for each grid cell; a value greater than 50 indicates that the model outperforms climatology. Different MOS models except for Raw in (a),(f),(k), which indicates the uncorrected S2S model output. In (top)–(bottom), the grid cells that observed a 90th percentile exceedance for 1–7 Dec 2015 are outlined in black.

Doss-Gollin et al (2018)

Outline

1. Model Output Statistics (MOS): bias correction, calibration.

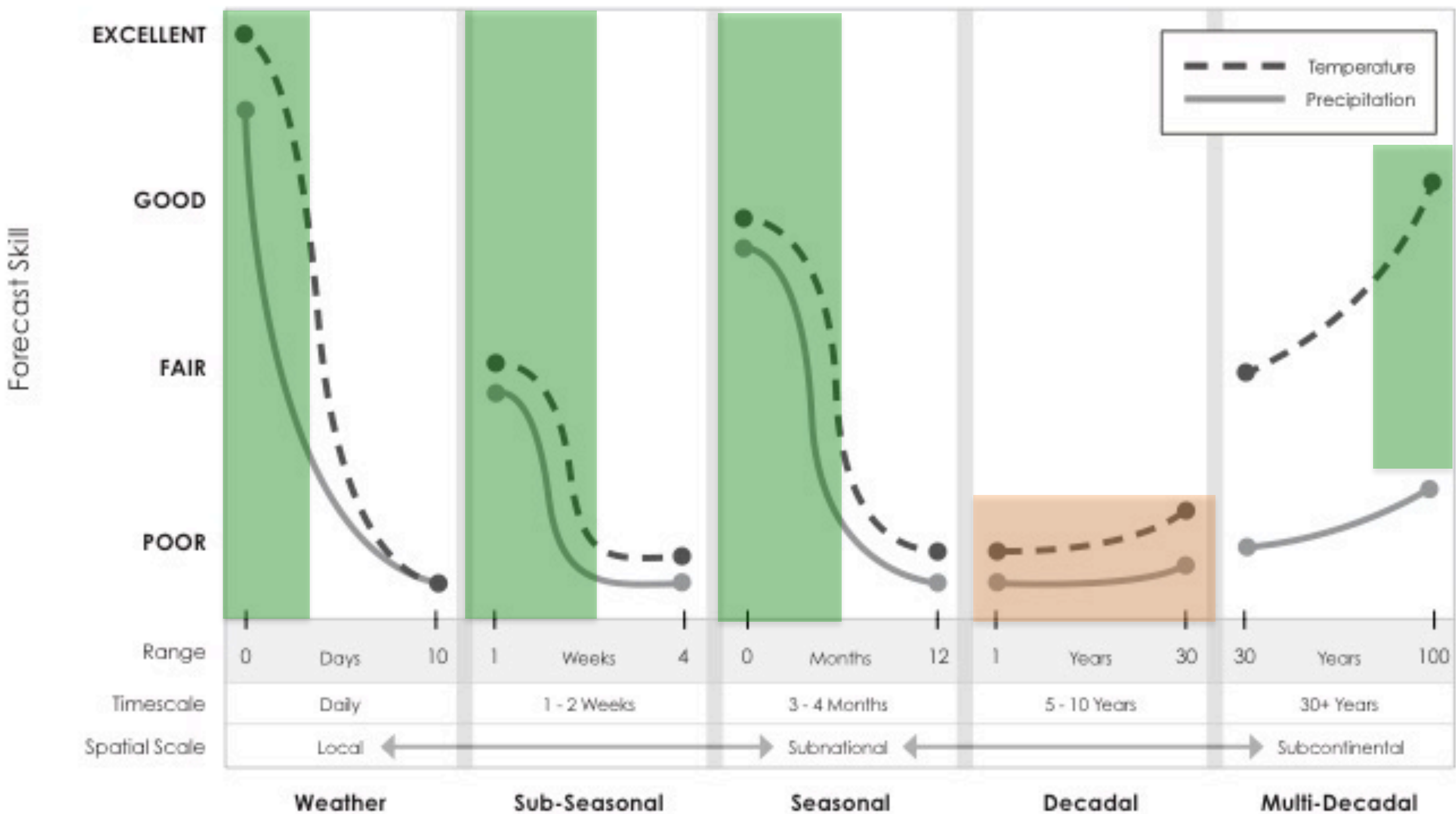
2. Local calibration techniques. Examples.

3. Non-local calibration techniques. Examples.

4. Exercises: pattern-based calibration with PyCPT

Before some hands-on examples with PyCPT, let's refresh some ideas…

Predictive skill depends on space and time scales

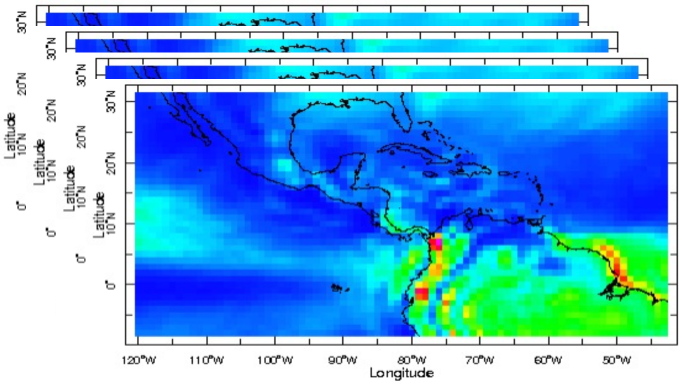


Thomson et al. (2018)

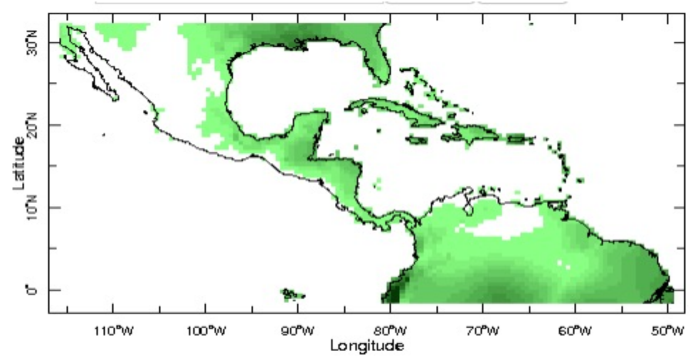
#NextGen: Multi-model calibration, ensemble and verification

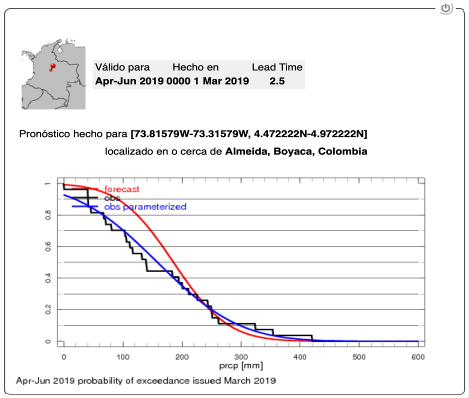
Raw Model Output (Predictors)

Multiple (C)GCMs [NMME, C3S, LC-LRFMME]

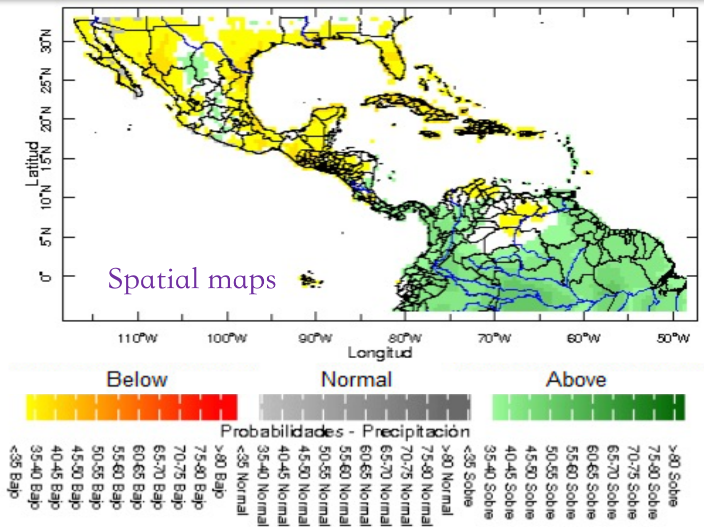


Gridded/Station Obs (Predictand)



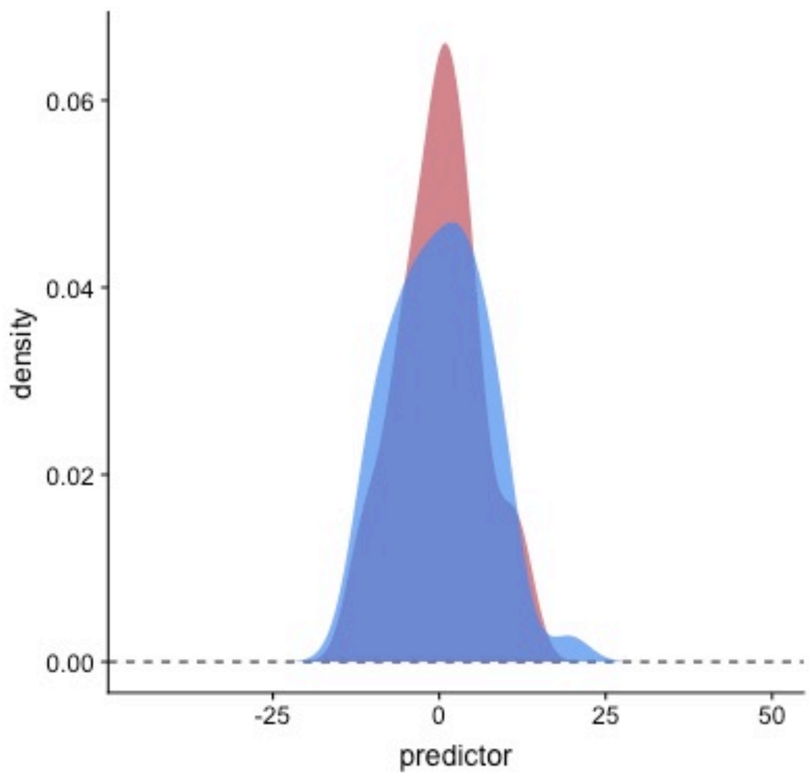


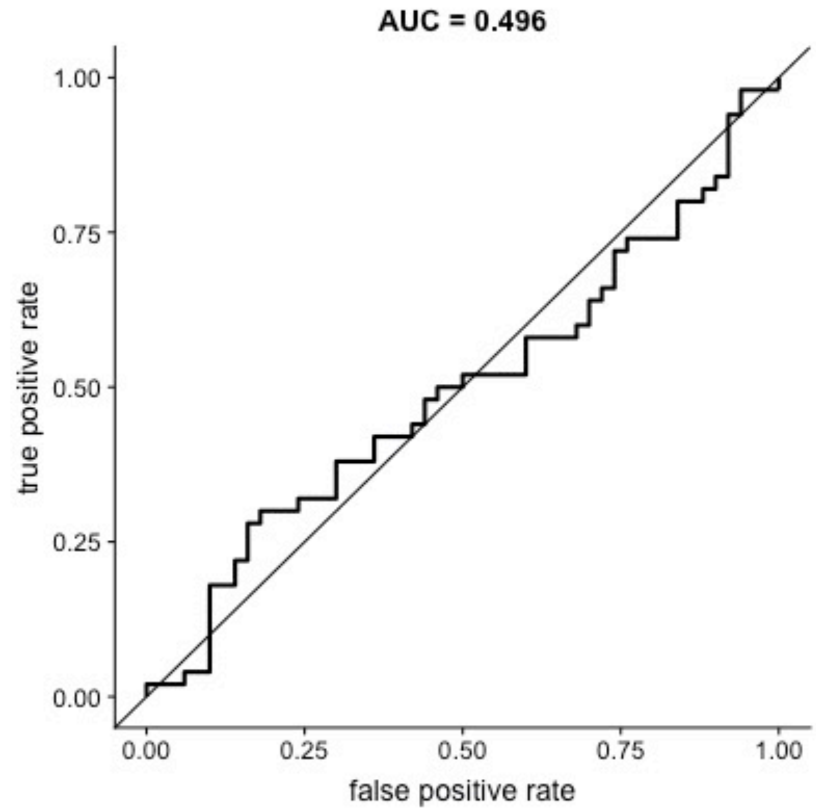
Flexformat! Use entire PDF



Pattern-based calibration (Model Output Statistics), each model independently, then ensemble in probability space

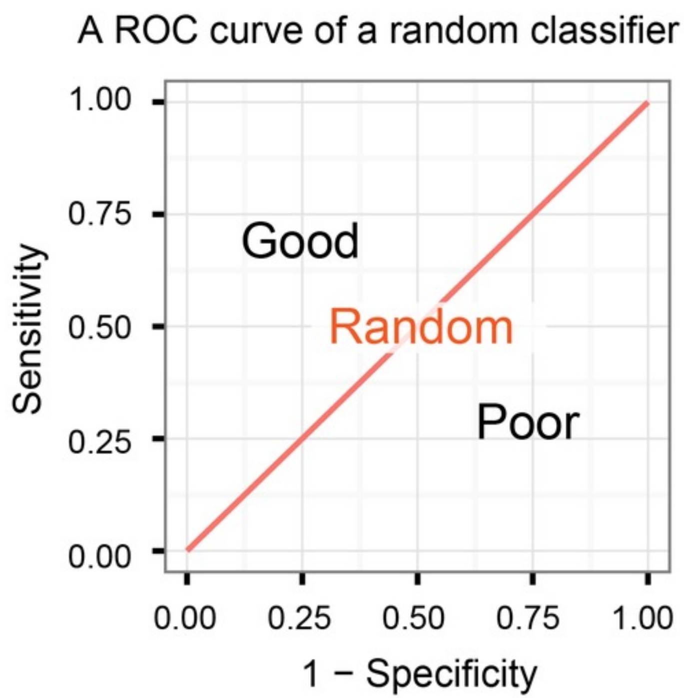
Discrimination – ROC



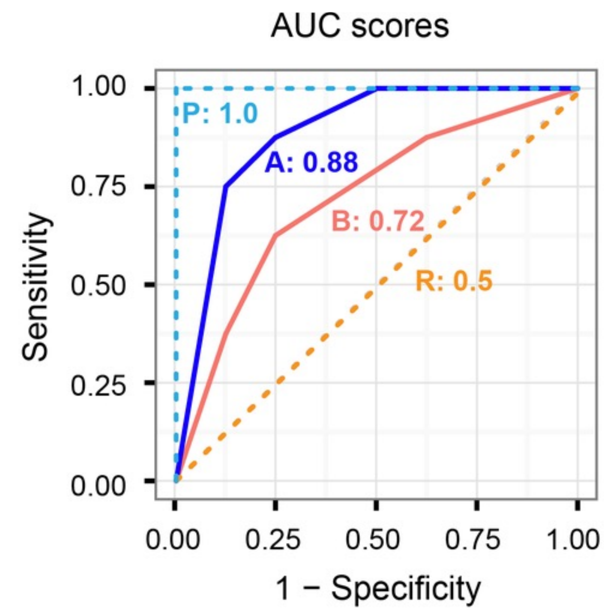


Courtesy of D. Sydykova

Discrimination – ROC

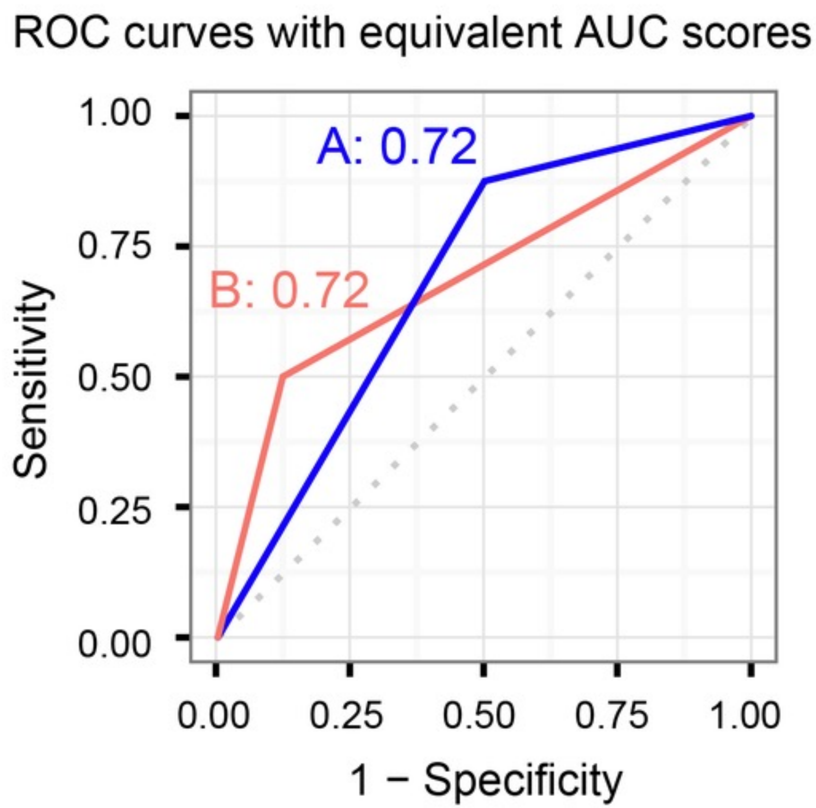


A ROC curve represents a classifier with the random performance level. The curve separates the space into two areas for good and poor performance levels.



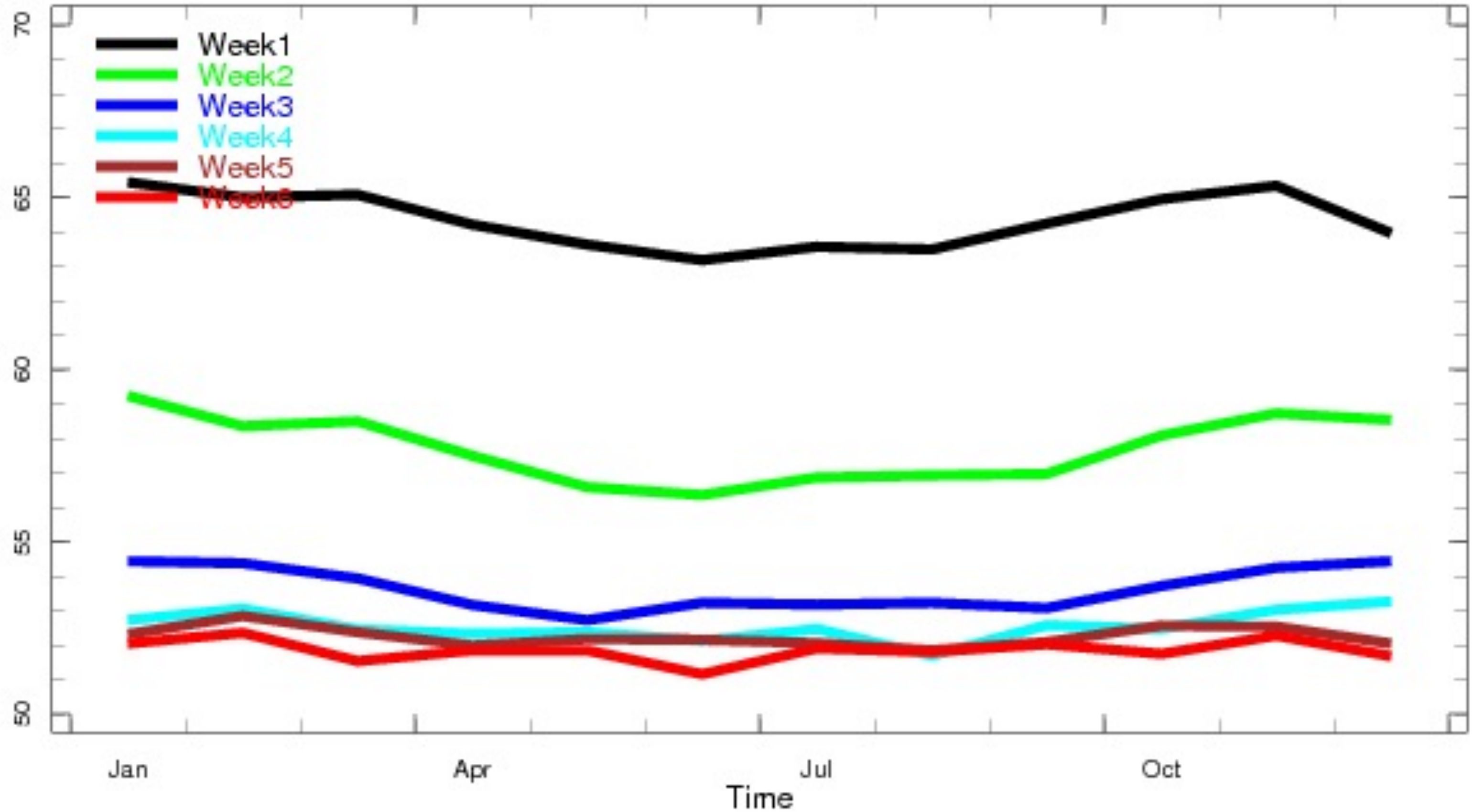
It shows four AUC scores. The score is 1.0 for the classifier with the perfect performance level (P) and 0.5 for the classifier with the random performance level (R). ROC curves clearly shows classifier A outperforms classifier B, which is also supported by their AUC scores (0.88 and 0.72).

Discrimination – ROC



Two classifiers A and B have the same AUC scores, but their ROC curves are different.

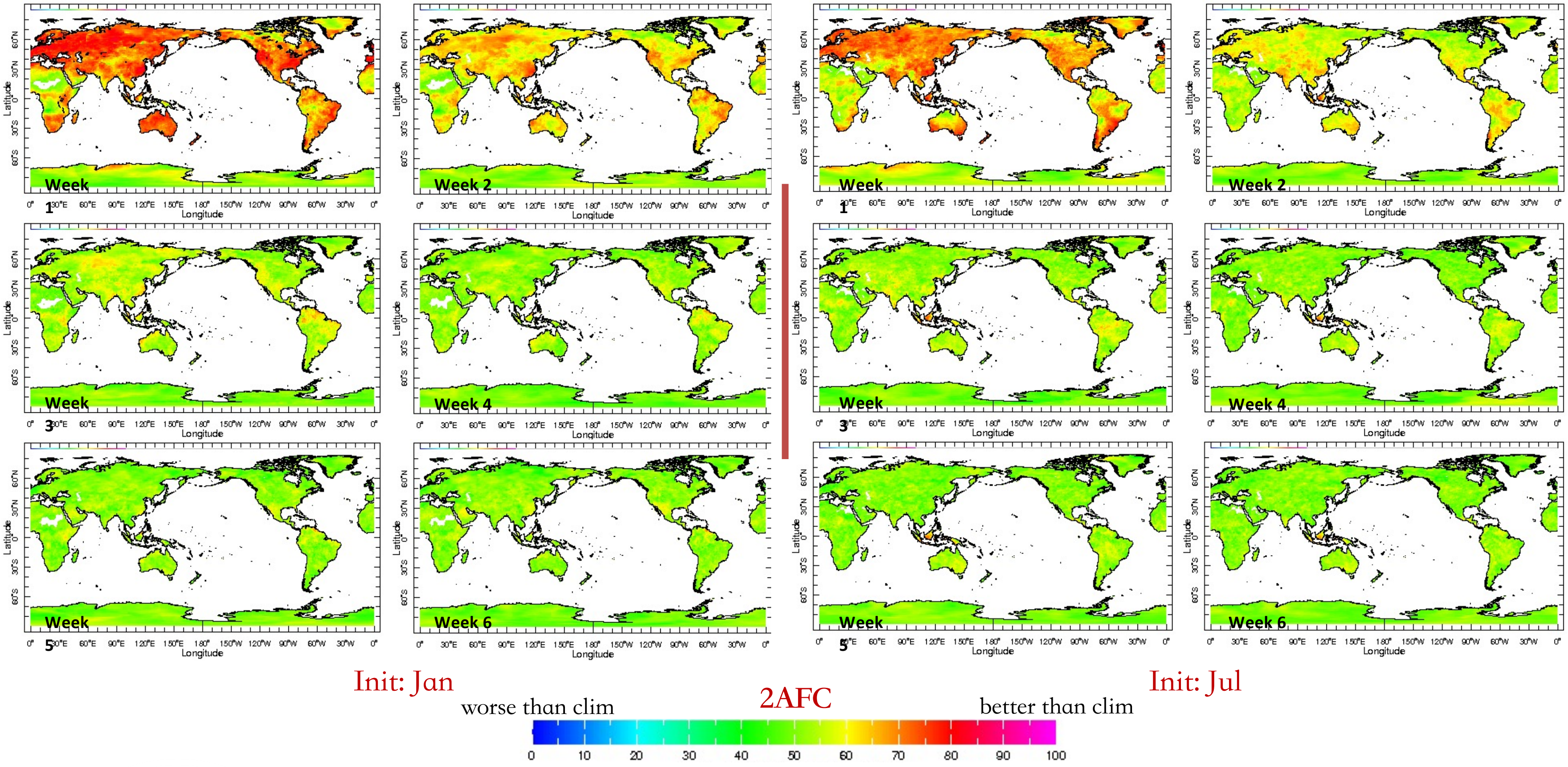
The Seasonality of Sub-Seasonal Skill



Global - Deterministic 2AFC - Model: ECMWF (uncorrected)

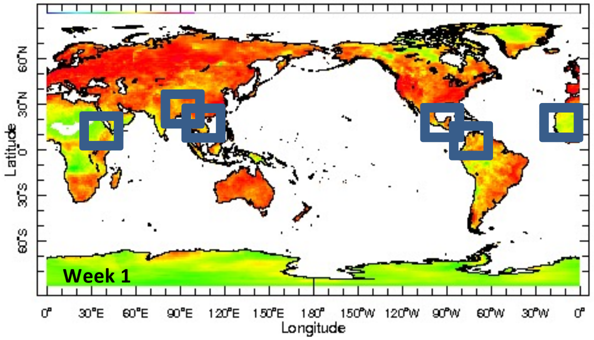
Muñoz *et al.* (2018)

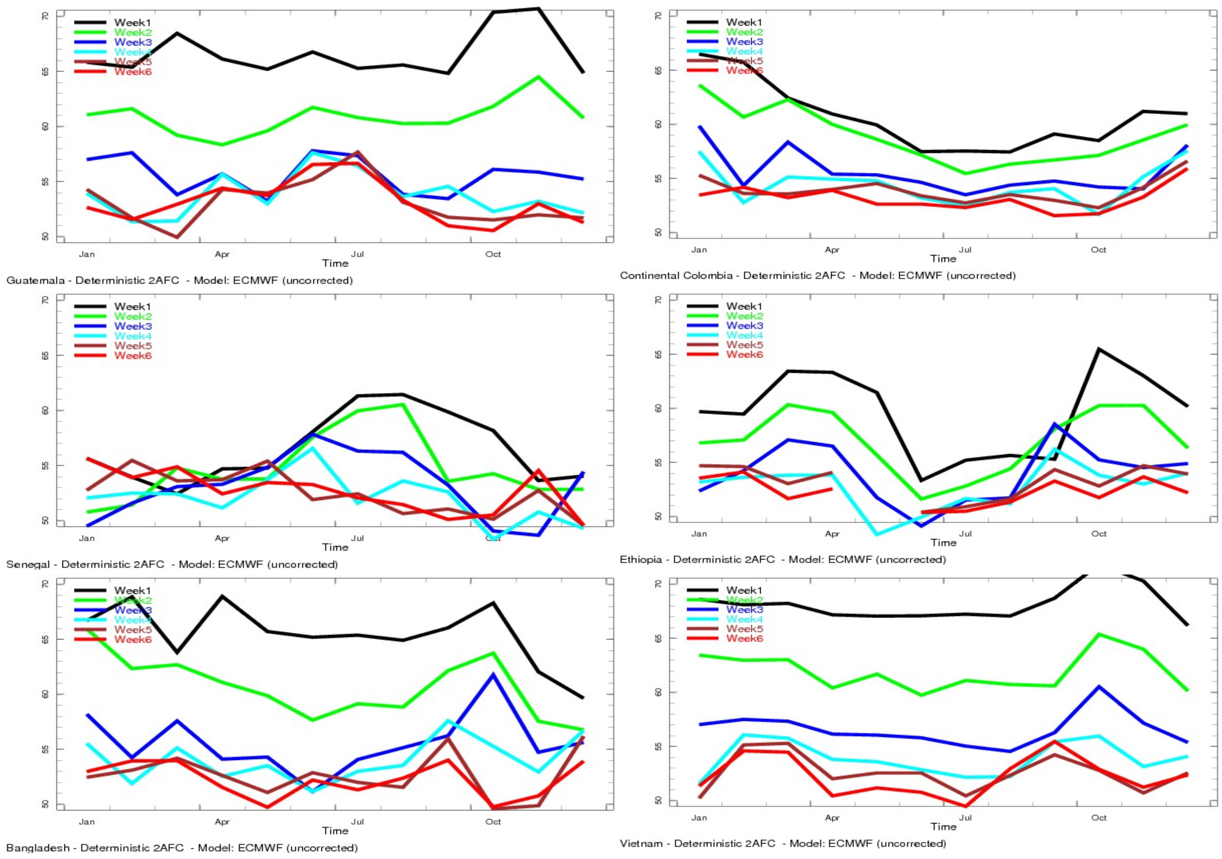
The Seasonality of Sub-Seasonal Skill



Muñoz et al. (2018)

The Seasonality of Sub-Seasonal Skill





2AFC

Muñoz et al. (2018)

Ignorance Score

The Ignorance Score (IGN), or negative log-likelihood score, of a probabilistic forecast of *n* categories can be written as (Good, 1952; Roulston & Smith, 2002):

$$IGN = -\log_2(p_k) \qquad\qquad k = 1..n$$

and it can be decomposed into reliability, resolution and uncertainty terms:

$$IGN = \underbrace{REL}_{\text{calibration}} - \underbrace{RES}_{\text{sharpness (if reliable)}} + \underbrace{UNC}_{\text{obs distribution}}$$

(Weijs et al., 2010; Wilks, 2018)

- It measures the information deficit, or ignorance, of a person having a probabilistic forecast but not knowing the actual outcome.
- Units are *bits* of information. IGN=0 means perfect forecast (zero ignorance).
- Each bit of ignorance represents a factor-of-2 increase in uncertainty.
- Related to expected gambling return if used to place proportional bets on the future (cost-loss scenarios).

Ignorance Score

The Ignorance Score (IGN), or negative log-likelihood score, of a probabilistic forecast of *n* categories can be written as (Good 1952; Roulston & Smith, 2002):

$$IGN = -\log_2(p_k) \qquad\qquad k = 1..n$$

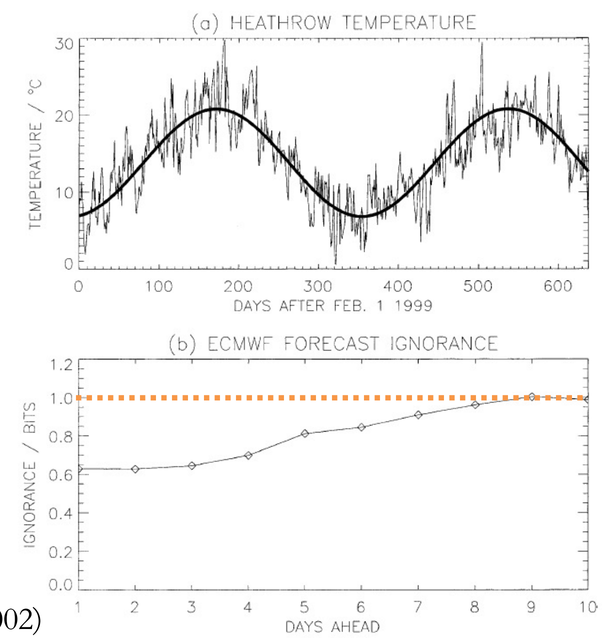
and it can be decomposed into reliability, resolution and uncertainty terms:

$$IGN = \underbrace{REL}_{\text{calibration}} - \underbrace{RES}_{\text{sharpness}} + \underbrace{UNC}_{\text{obs distri}}$$

(Weijs et al., 2010; Wilks, 2018)

There are different ways to define a skill score for IGN. Here we use climatology as the reference. For equiprobable climatological categories,

$$ISS = -\frac{\log_2(p_k)}{\log_2(n)} \begin{cases} > 1 & \text{Less info than climatology} \\ = 1 & \text{As good as climatology} \\ < 1 & \text{More info than climatology} \end{cases}$$

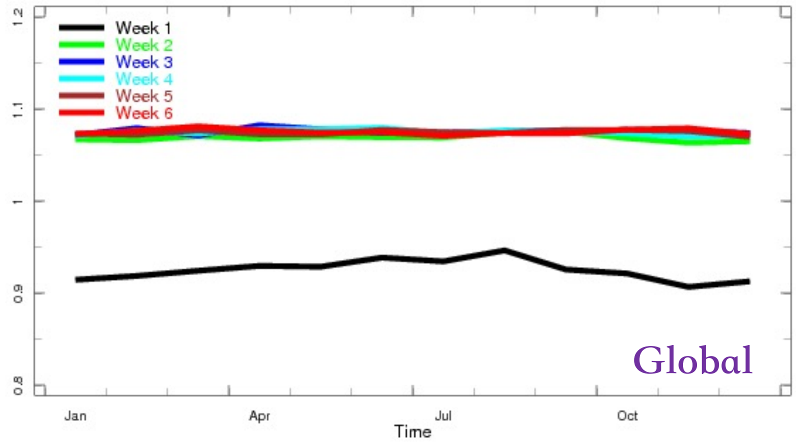


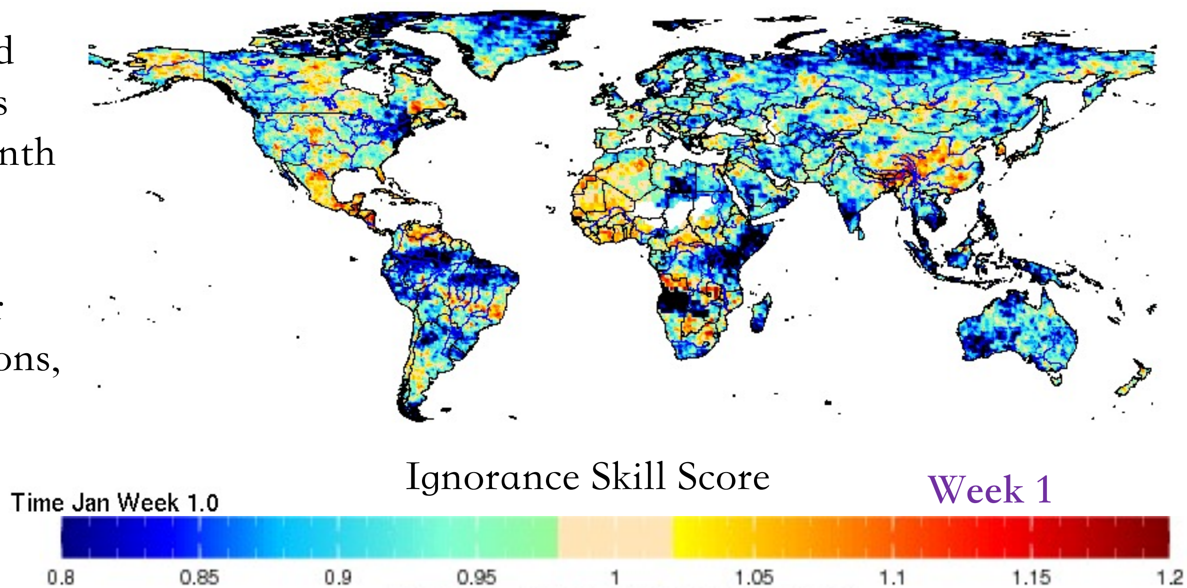
Roulston & Smith (2002)

Seasonality of Sub-seasonal Skill

<u>Skill Assessment</u>

- Model: ECMWF
- Rainfall
- Probabilistic Hindcasts
- Obs: CPC Unified
- All initializations available per month (8-9)
- **Uncalibrated**
- IGN, RPSS, Brier and decompositions, for Week 1-6

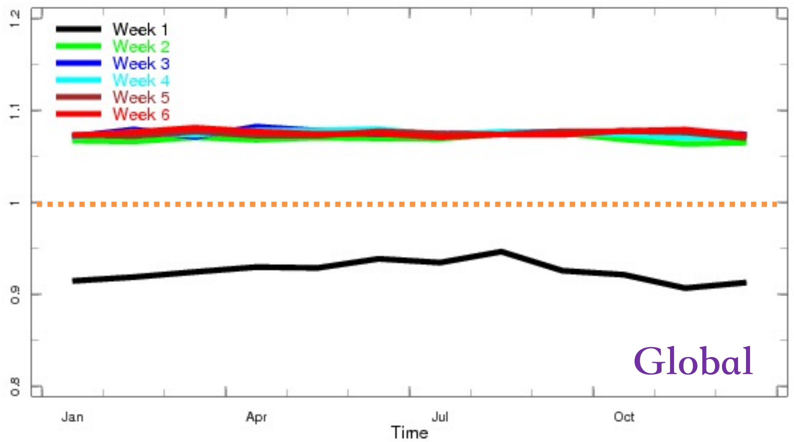


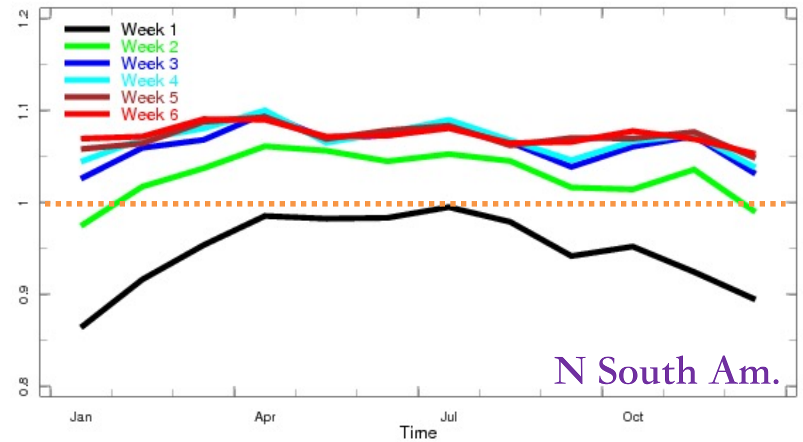


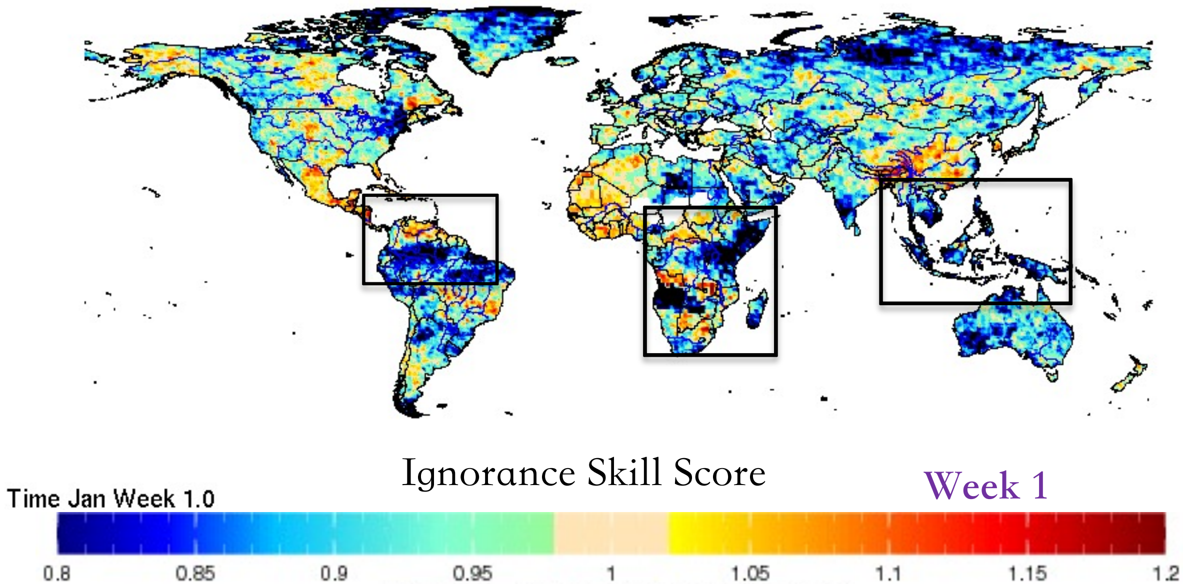
Muñoz et al. (2018)

Better Climatology Worse

Seasonality of Sub-seasonal Skill







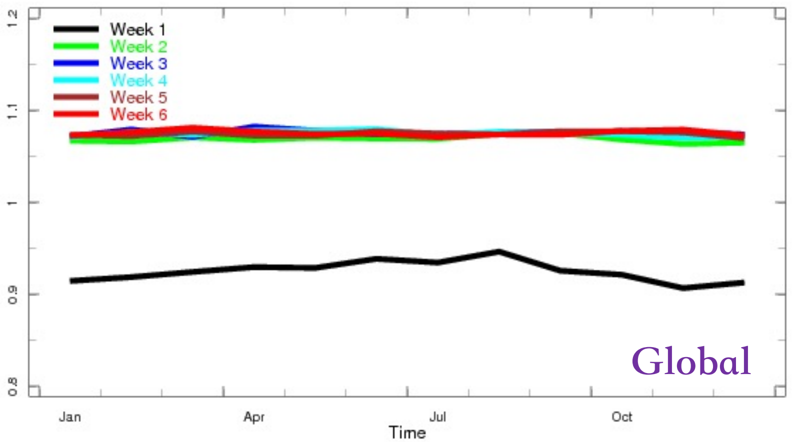
Ignorance Skill Score

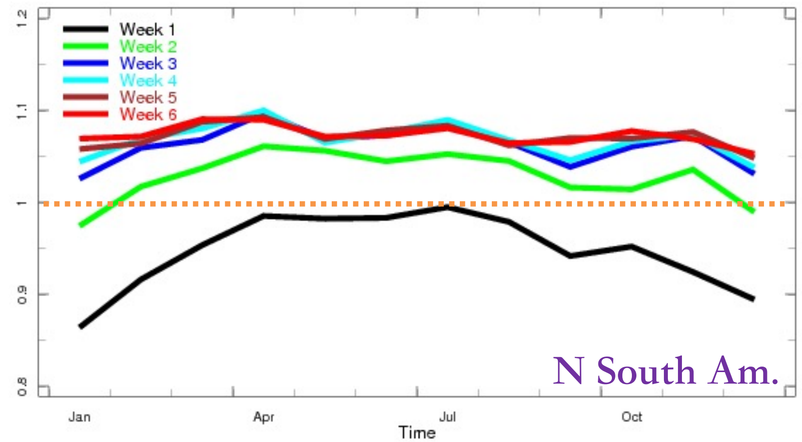
Week 1

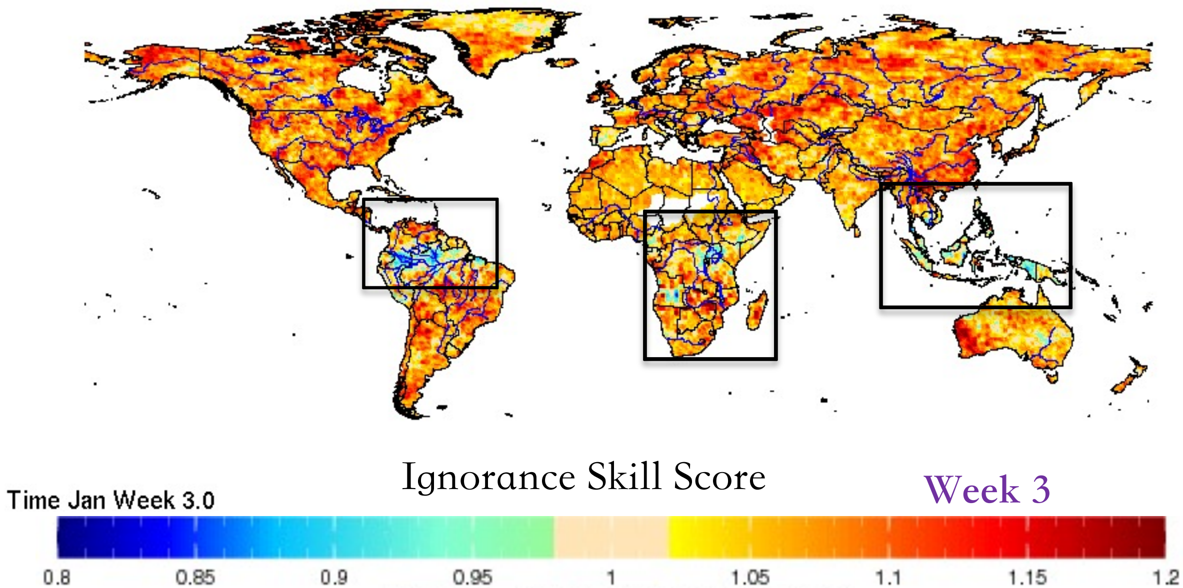
Muñoz et al. (2018)

Better Climatology Worse

Seasonality of Sub-seasonal Skill



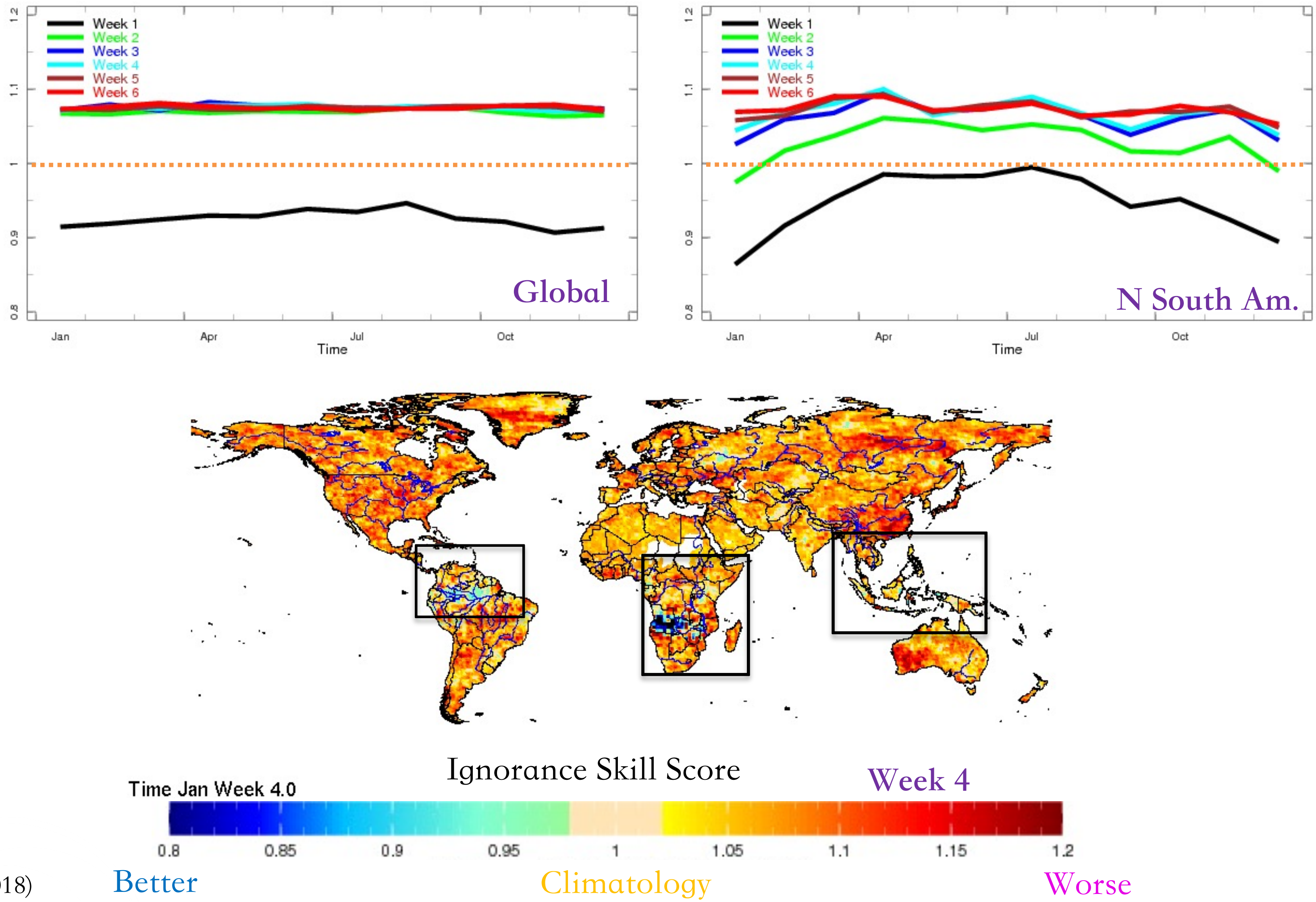




Muñoz et al. (2018)

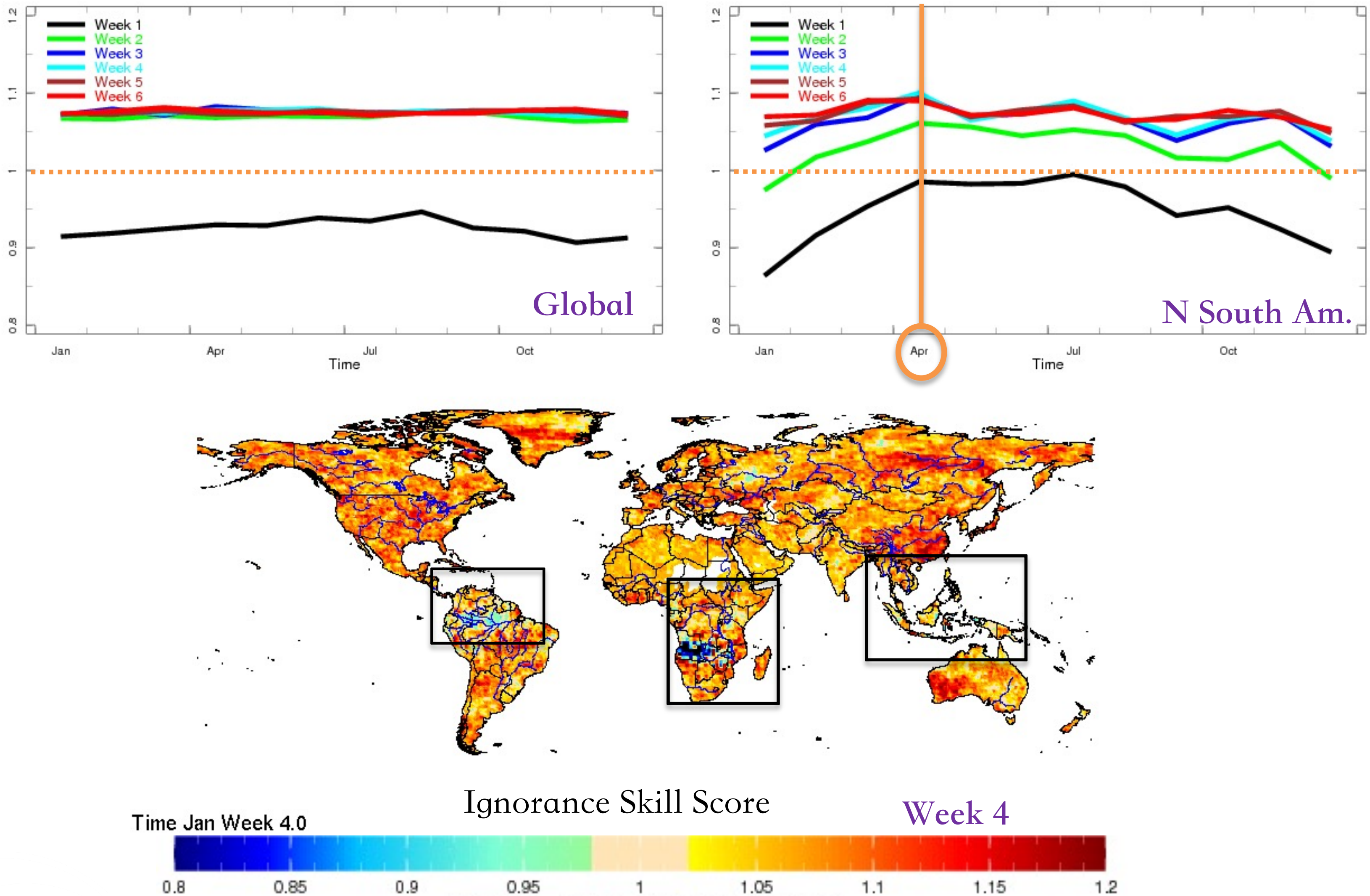
Better Climatology Worse

Seasonality of Sub-seasonal Skill



Muñoz et al. (2018)

Seasonality of Sub-seasonal Skill



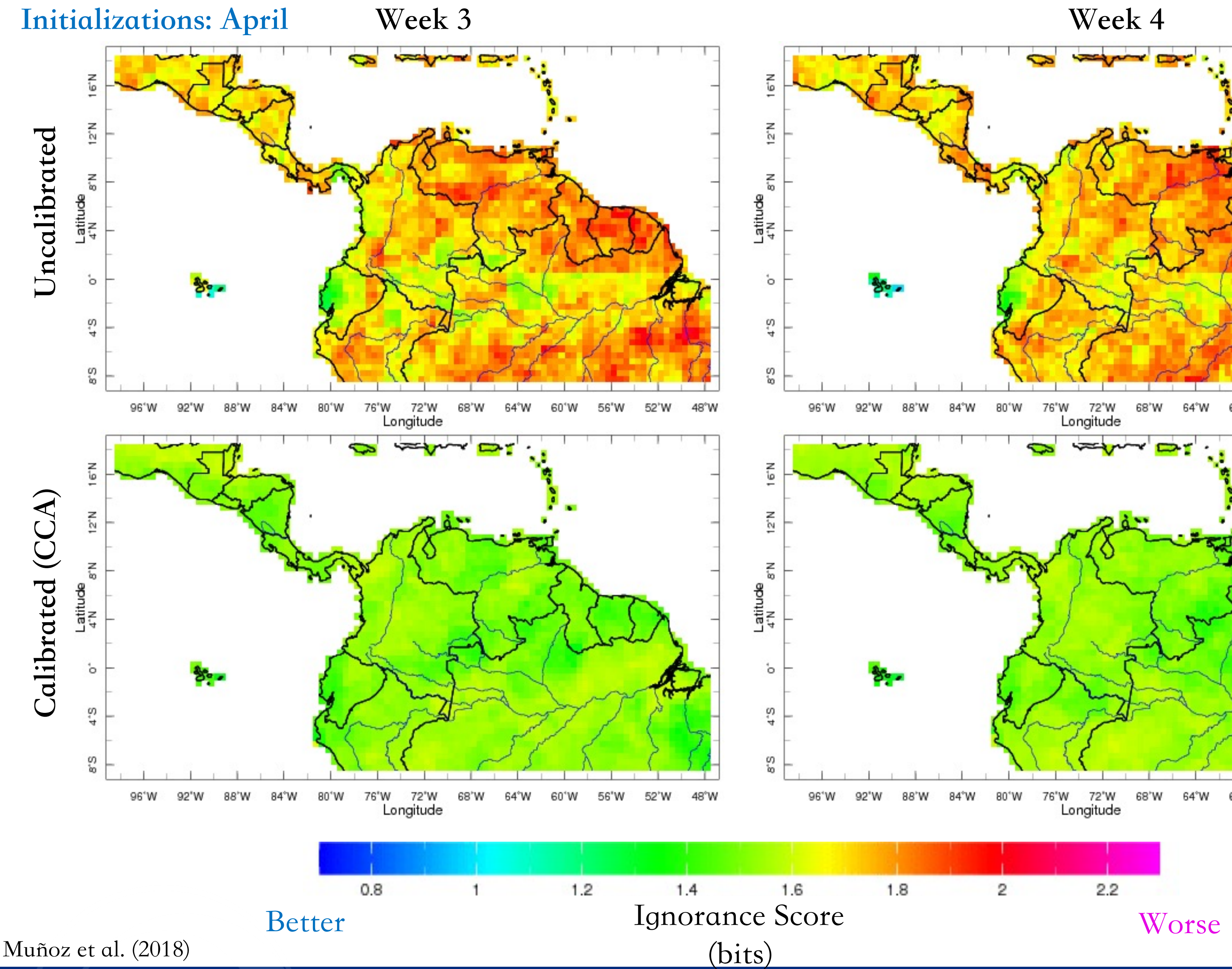
Muñoz et al. (2018)

Better

Climatology

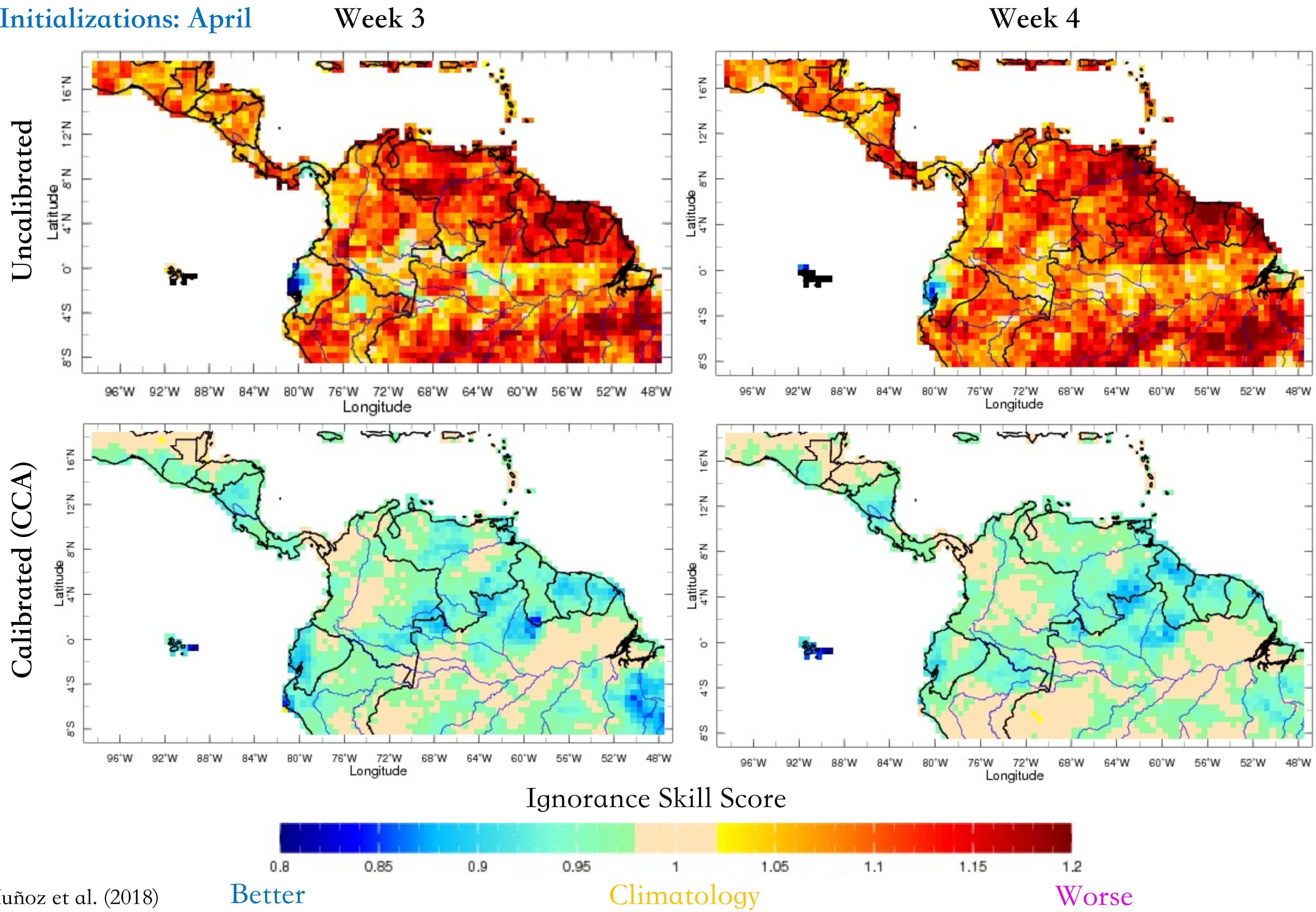
Worse

Initializations: April



Muñoz et al. (2018)

Initializations: April



Muñoz et al. (2018)

Let's PyCPT!

Conclusions

- Generally speaking, we *always* need to calibrate (MOS) our raw forecasts.
- Multiple techniques, depending on the desired outcome (different forecast attributes).
- Improving one particular skill metric can decrease skill in other metrics.
- Local and non-local (or pattern-based) calibrations have their own pros and cons.
- Model Output Statistics has the potential to improve forecast skill at sub-seasonal timescales. In particular, EOF-based MOS methods like Canonical Correlation Analysis (and Principal Component Regression) show clear skill improvement for different regions of the world, both in magnitudes and spatial patterns, <u>but not always</u>.

Local versus non-local calibration techniques

Ángel G. Muñoz

with contributions from
Simon J. Mason
Andrew W. Robertson

International Research Institute
for Climate and Society
Earth Institute | Columbia University