Seasonal Forecast Verification Significance & Robustness (Interpretation of scores)

JP. Céron
and contribution from S. Mason (IRI)
jpceron.wmo@gmail.com

WEATHER CLIMATE WATER
TEMPS CLIMAT EAU

WMO OMM

World Meteorological Organization
Organisation météorologique mondiale

lectures

1. *Introduction*
2. *Forecast Attributes*
3. *ROC & Reliability - Exercise*
4. **Significance & Robustness**
5.

Introduction

Given a verification score how do we know whether the score's value is good, especially if the score is some abstract number like the Spearman's correlation that has no obvious interpretation?

*By a **"good" score** we could mean:*

1. *It indicates that the forecasts are skilful.*
2. ***It gives a realistic estimate of how good the forecasts are.***

The second point is notably related to the question of the distribution of the score (pdf) and its confidence interval (under the hypothesis that there is a forecasting information potentially « good » or not).

It's a fundamental question to adress before implementing a forecasting model for routinely provide forecasts ; it's related to the robustness of the performance of the implemented model.

Introduction

Generally speaking, the score is related to the influence of the forecasted information X (e.g. predictors) on the targetted information Y (e.g. predictand). It is related to the relationship (e.g. mostly statistical in downscaling) infered through the observed dataset with respect of the choosen model (whatever statistical or dynamical).

So, looking to a causal relationship Φ *between X and Y, the calibration of the model will infer a function* $\tilde{\Phi}$ *which take into account this relationship*

$$\tilde{Y} = \tilde{\Phi}(X)$$

However, the $\tilde{\Phi}$ *relationship is only an estimation of the true function* Φ

So we can write that

$\tilde{\Phi} = \Phi + \hat{\Phi}$ *where* $\hat{\Phi}$ *depends on sampling and method errors*

the $\hat{\Phi}$ *component gives non reproducible scores while the component gives the reproducible part and should be estimated as well as possible*

Introduction

In downscaling when estimating scores :

$\hat{\Phi}$ *generally leads to statistical instabilities*

$\tilde{\Phi}$ *can leads to numerical instabilities (in case of multicolinearity within the predictor dataset - e.g. impossibility to compute the inverse Variance-Covariance matrix in MR or LDA)*

The real score corresponding to Φ *is the good answer to point 2*

However generally we don't know any thing about the Φ *function*

So using scores computed on the samples, that is to say using $\tilde{\Phi}$ *we must try to* ***infer the pdf of the real score*** *and then* ***answer to the question 2***

Introduction

To address sampling uncertainties

Test file independent from the learning file ; but partial answer to the point 2

Cross-validation (notably when sampling size is limited – same limitation than test file – partial answer)

Test files ; better answer because of the knowledge of the distribution (but generally sampling size limitation)

Bootstrapping (re-sampling method, a way to get “several test files”)

Randomisation of the files (trying to estimate $\hat{\phi}$ and associated score)

To address model uncertainties

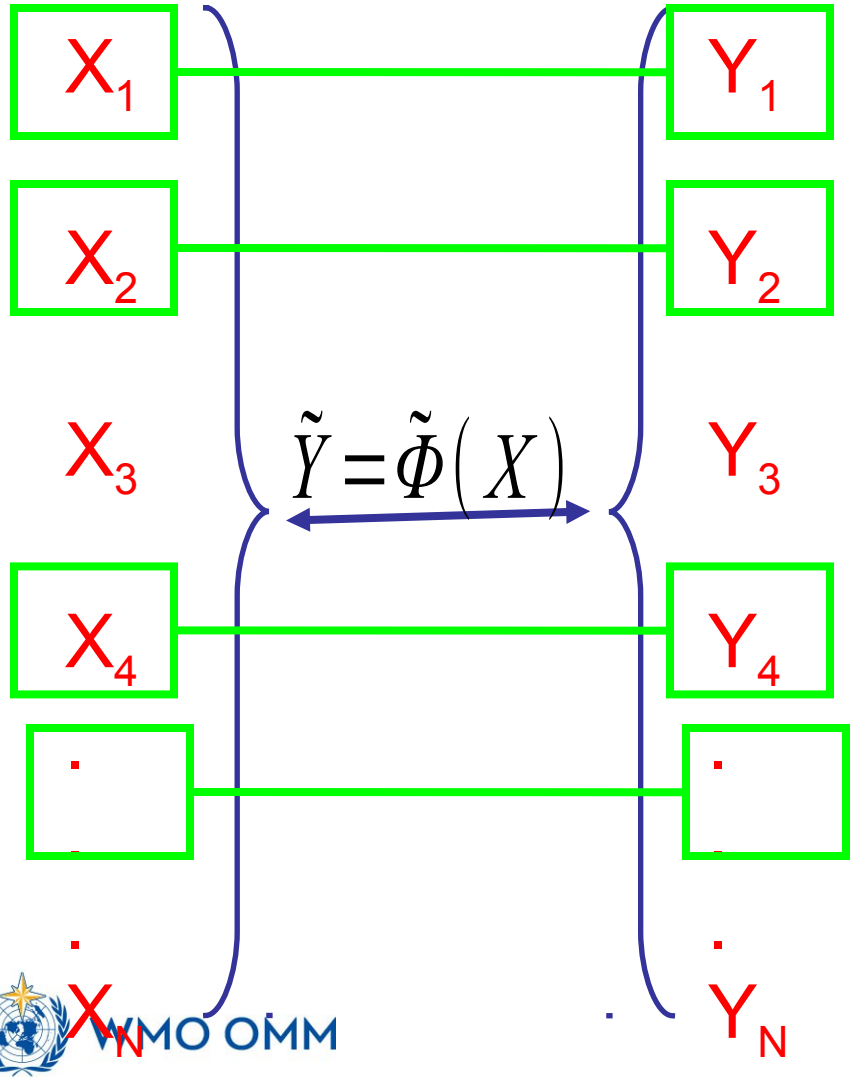
Multi-model approach (similar to several test files)

To address uncertainties

Monte-Carlo methods (random perturbations allowing to issue several samples e.g. perturbing the slope of the regression)

Introduction

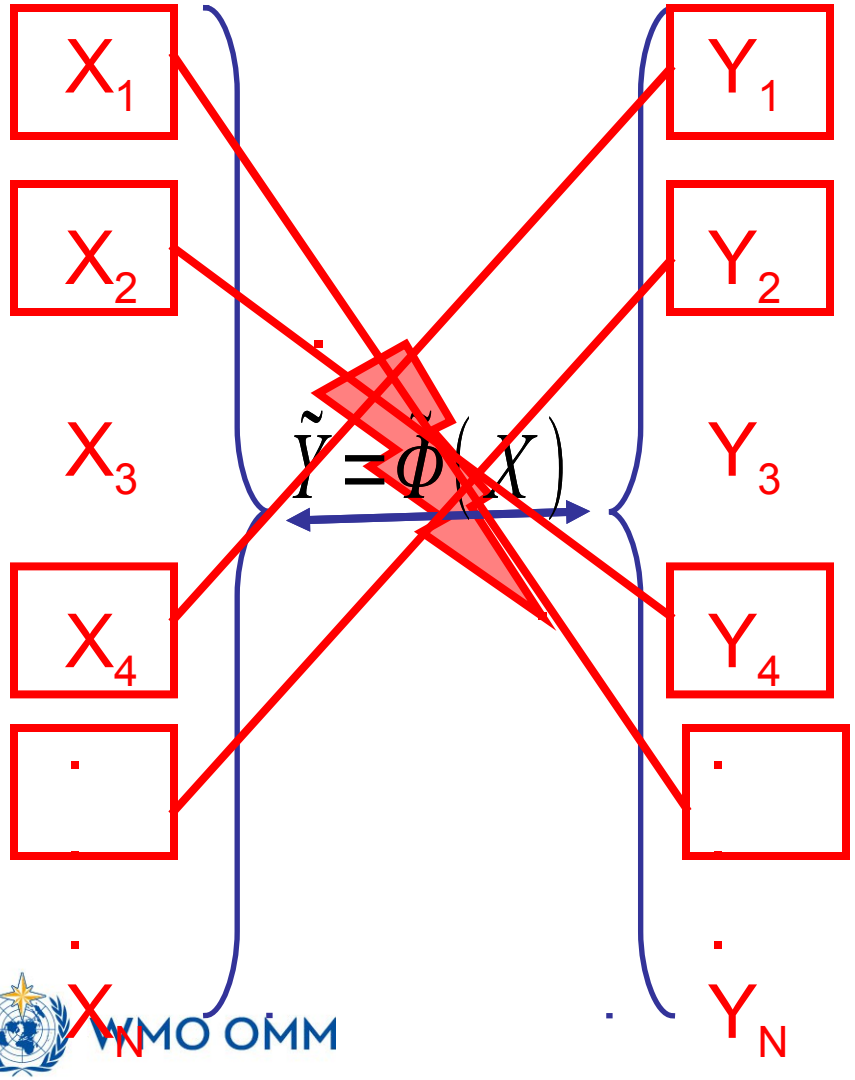
- **The main idea : resampling vs randomisation**



- The $\tilde{\Phi}(X)$ is induced by each pair (X_i, Y_i) for $i = 1,N$ (synchroneous aspect)
- In a **resampling procedure**, one try to get several samples preserving the "synchroneous" aspect (subset of the initial dataset) in order to have some insight into the distribution of the score

Introduction

- **The main idea : resampling vs randomisation**



- The $\tilde{\Phi}(X)$ is induced by each pair (X_i, Y_i) for $i = 1,N$ (synchroneous aspect)
- In a **randomisation process** one try to get several samples breaking the synchroneous aspects in order to estimate the distribution of the score wihout any relationship

P-values

Given a verification score how do we know whether the score's value is good, especially if the score is some abstract number like the Spearman's correlation that has no obvious interpretation?

A commonly used method to assess whether a verification score's value is "good" is to calculate the probability that a value at least as good as that observed could have been achieved given completely useless forecasts.

For instance the M^2 (squared multiple correlation) or Δ^2 (Mahalanobis distance) can reach "high" values without any useful information

$$\bar{M}^2 \approx pN - 1$$
$$\bar{\Delta}^2 \approx pN_A + pN_{NA}$$

Under the H0 hypothesis of No Information (No correlation or No discrimination)

This probability is called a *p*-value.

P-values

Calculating p-values :

all methods involve defining a distribution of scores under the null hypothesis of no skill. There are a number of ways of obtaining this distribution:

1. ***Exact theoretical distribution*** **:** e.g., binomial for hit rates, *U* for ROC area
2. ***Approximate theoretical distribution*** : e.g., Student's *t* for correlation, Gaussian for ROC area.
3. ***Empirical distribution*** : using permutation methods (randomisation of the files).
4. ***Empirical distribution*** **:** using artificial series (Monte-Carlo methods).

P-values

3. Approximate the distribution by generating a large number of random rankings. Numerous samples are issued (typically a few hundred up to a thousand)

   A randomisation of the initial file is done using *permutation* procedure to obtain the random rankings

Year	Obs.	Year	For.
2001	-0.23	2001	0.28
2002	1.59	2002	0.77
2003	0.41	2003	0.44
2004	0.92	2004	0.59
2005	-0.55	2005	0.37

Permutation 1

Year	Obs.	Year	For.
2001	-0.23	2005	0.37
2002	1.59	2004	0.59
2003	0.41	2001	0.28
2004	0.92	2003	0.44
2005	-0.55	2002	0.77

Permutation 2

Year	Obs.	Year	For.
2001	-0.23	2004	0.59
2002	1.59	2001	0.28
2003	0.41	2003	0.44
2004	0.92	2005	0.37
2005	-0.55	2002	0.77

Permutation 3

Year	Obs.	Year	For.
2001	-0.23	2003	0.44
2002	1.59	2001	0.28
2003	0.41	2005	0.37
2004	0.92	2002	0.77
2005	-0.55	2004	0.59

P-values

P-values indicate only ***how confident we can be that our forecasts have some skill****; the actual amount of skill that we may have could be exceedingly small.*

So a small *p*-value (e.g. below 0,05) allows us only to say:

"I am **very confident** that **I do not have No Skill**."

A large *p*-value (e.g. above 0,80) highlight that the forecast score, whatever its value, can be obtain with purely random data.

In that case, it is highly recommended to check back all the model component to diagnose what could be wrong (e.g. too much predictors vs number of cases)

P_Values

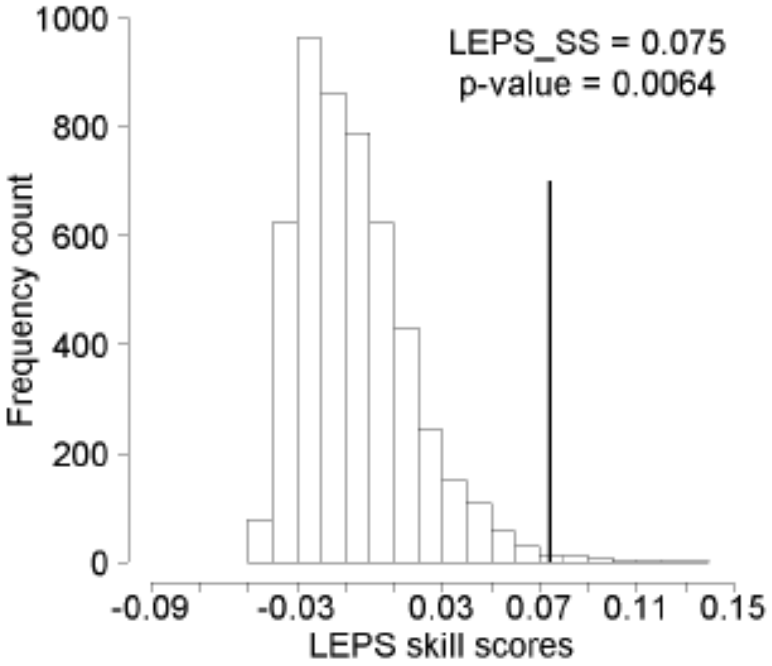


Fig. 1. Empirical null distribution for tercile LEPS skill score arising from the SOI forecast system for predicting JJA rainfall at Dalby (North-eastern Australia). The dark, thick line indicates the location of the observed LEPS_SS value.

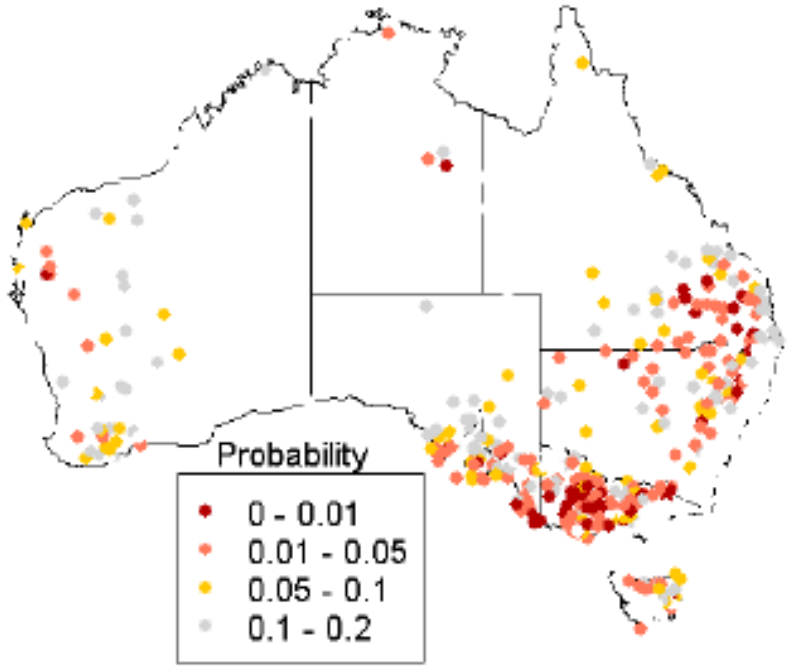


Fig. 2. Discriminatory ability of 5-phase SOI system for JJA rainfall across Australia, as measured by Kruskal-Wallis p-values. This test accounts for differences among class medians.

Confidence intervals

So, if ***"I am very confident that I do not have no skill", how much skill do I have?***

Example : I got a ***correlation coefficient of 0.5*** *on a* ***sample of 30 observations*** *what's about the real value ?*

Values so different as ***0,2 or 0,7*** *can be consistent with my score (I_{95} ~ [0.16 , 0.73])*

The sample score provides one indication of the skill. But is this value correct?

Problem *: you have* ***only ONE value*** *which is not necessarily representative of all possible values (pdf of the score)*

A good solution : several independent test files but … needs of very large size dataset ?

Confidence intervals

If we had a different set of forecasts the calculated score will vary from the sample score even if the skill of the forecasts is unchanged.

*It would be helpful to know **how sensitive the score is to the sample** ; if the score is sensitive the uncertainty in the estimate will be high..*

*A recommended way of indicating uncertainty is to **calculate confidence intervals of the score** using different samples.*

Problem : a large number of samples is needed ; how to cope with this need ?

*Among the **different factors of uncertainty** on the score one can point out : the size of the sample, the method used, the predictors themselves and also to the data quality*

Confidence intervals

Why calculate confidence intervals?

1. *Indicates sampling uncertainty in the score.*
2. *More informative than p-values.*
3. *Facilitates comparison of scores.*

Confidence intervals

There are many ways of calculating confidence intervals. Some of the most commonly used procedures include:

1. ***Exact theoretical distribution*:** e.g., binomial for hit rates,
2. ***Approximate theoretical distribution***: e.g., Student's *t* for ROC area.
3. ***Empirical distribution*** : using bootstrap methods (or several test files).

Confidence intervals

A *bootstrap* procedure is commonly used to obtain the resamples

A *bootstrap* procedure involves **resampling with replacement** (compare with the permutation procedure in which the object is to generate useless sets of forecasts).

The size of the bootstrapped samples should be consistent with the initial sample (e.g. from 80% to 90% of the initial size)

Numerous samples are issued

(typically a few hundred up to a thousand)

Confidence intervals

A *bootstrap* procedure is used to obtain the resamples (compare with the permutation procedure)

Year	Obs.	Year	For.
2001	-0.23	2001	0.28
2002	1.59	2002	0.77
2003	0.41	2003	0.44
2004	0.92	2004	0.59
2005	-0.55	2005	0.37

Bootstrap 1

Year	Obs.	Year	For.
2001	-0.23	2001	0.28
2001	-0.23	2001	0.28
2002	1.59	2002	0.77
2004	0.92	2004	0.59
2005	-0.55	2005	0.37

Bootstrap 2

Year	Obs.	Year	For.
2002	1.59	2001	0.28
2002	1.59	2001	0.28
2003	0.41	2003	0.44
2003	0.41	2003	0.44
2004	0.92	2005	0.37

Bootstrap 3

Year	Obs.	Year	For.
2002	1.59	2001	0.28
2002	1.59	2001	0.28
2002	1.59	2001	0.28
2004	0.92	2002	0.77
2005	-0.55	2004	0.59

Confidence intervals

Sample

No.	Obs 1	Obs 2	No.	For 1	For 2
1	-0.08	-0.23	1	0.28	0.16
2	1.56	1.59	2	0.77	0.87
3	0.58	0.41	3	0.44	0.34
4	0.90	0.92	4	0.59	0.71
5	-0.21	-0.55	5	0.37	0.19

Bootstrap 1

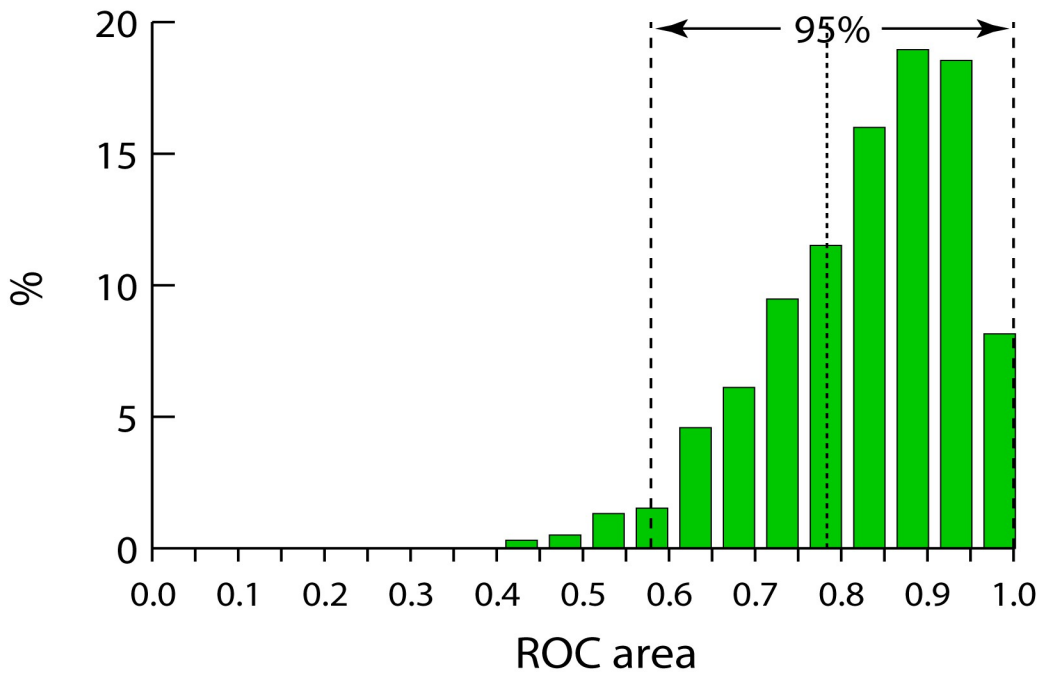
No.	Obs 1	Obs 2	No.	For 1	For 2
1	-0.08	-0.23	1	0.28	0.16
1	-0.08	-0.23	1	0.28	0.16
3	0.58	0.41	3	0.44	0.34
4	0.90	0.92	4	0.59	0.71
4	0.90	0.92	4	0.59	0.71

Permutation 2

No.	Obs 1	Obs 2	No.	For 1	For 2
2	1.56	1.59	5	0.37	0.19
2	1.56	1.59	5	0.37	0.19
2	1.56	1.59	5	0.37	0.19
4	0.90	0.92	4	0.59	0.71
5	-0.21	-0.55	5	0.37	0.19

A *bootstrap* procedure is used to obtain the resamples (compare with the permutation procedure)

Confidence intervals



Note:

1. the sample score can be biased (some bootstrap procedures adjust for this);
2. the distribution of skill scores generally will be skewed.

Cross-Validation

Model fitting does not provide a good estimate of actual forecast skill: the model-fit statistics tell us how well the model describes the data, not how well it predicts the data. We can describe the data perfectly by having enough variables in the model, but obviously this would not guarantee perfect forecasts.

Cross-Validation

To estimate true predictive skill we need a set of forecasts that are independent of the data used to train the model.

Specifically, the verification sample should be completely distinct from the training sample.

Any "leakage" of information from the training sample to the verification sample will bias the predictive skill estimate.

Cross-Validation

Cross validation (without a hyphen) means angry validation.

Cross-validation (with a hyphen) is a commonly used method for assessing how good a set of predictions are. It attempts to address the problem of obtaining a realistic estimate of the quality of the forecasts.

Cross-Validation

General procedure:

- Leave at least one year out of the training sample.
- Reconstruct the model using the new smaller training sample.
- Forecast at least one of the years omitted.
- Repeat at least step 3.

Objective: **mimic the complete lack of knowledge of future values in operational forecasting.**

Cross-validation

Leave-k-out cross-validation:

- Leave *k* contiguous years out of the training sample.
- Reconstruct the model using the new smaller training sample.
- Forecast the middle one of the omitted years.
- Repeat until a forecast has been made for each year.

Choice of K :

- Large enough to ensure the independancy of the taining file and the forecast
- Small enough to have enough forecasts and training files
- ~5 is a « good » compromise in seasonal forecast

Leave-*k*-out cross-validation

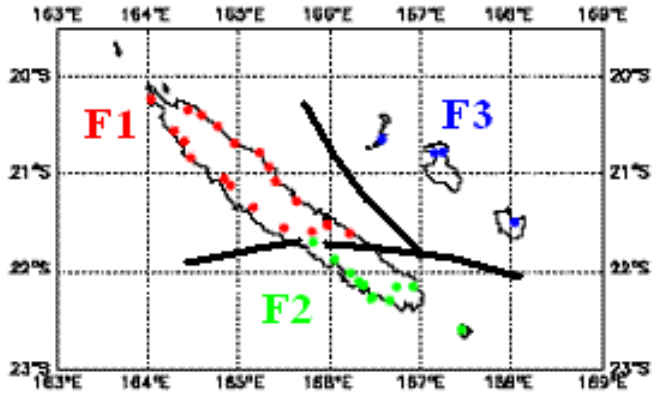
1951	Predict 1951	Omit 1952	Omit 1953	Training period			
1952	Omit 1951	Predict 1952	Omit 1953	Omit 1954	Training period		
1953	Omit 1951	Omit 1952	Predict 1953	Omit 1954	Omit 1955	Training period	
1954	Training period	Omit 1952	Omit 1953	Predict 1954	Omit 1955	Omit 1956	Training period
1955	Training period		Omit 1953	Omit 1954	Predict 1955	Omit 1956	Omit 1957

Leaving out more than one year reduces the negative bias, although too many years out causes large sampling errors.

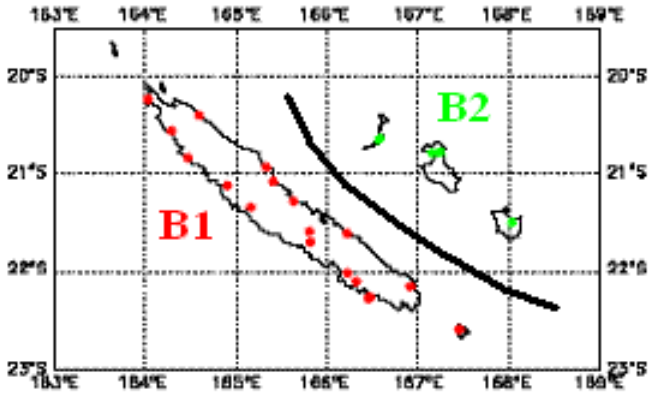
Downscaling

- *Downscaling in Space - Zoning*

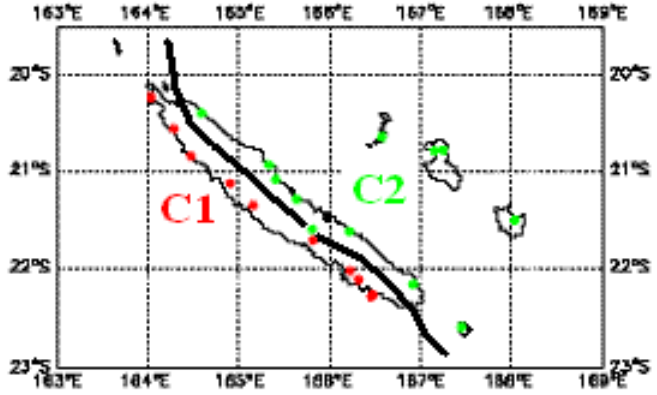
a)

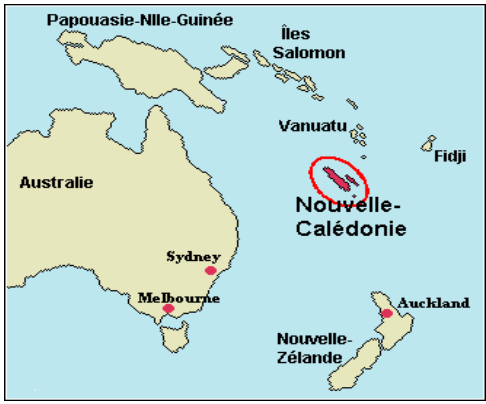


b)



c)





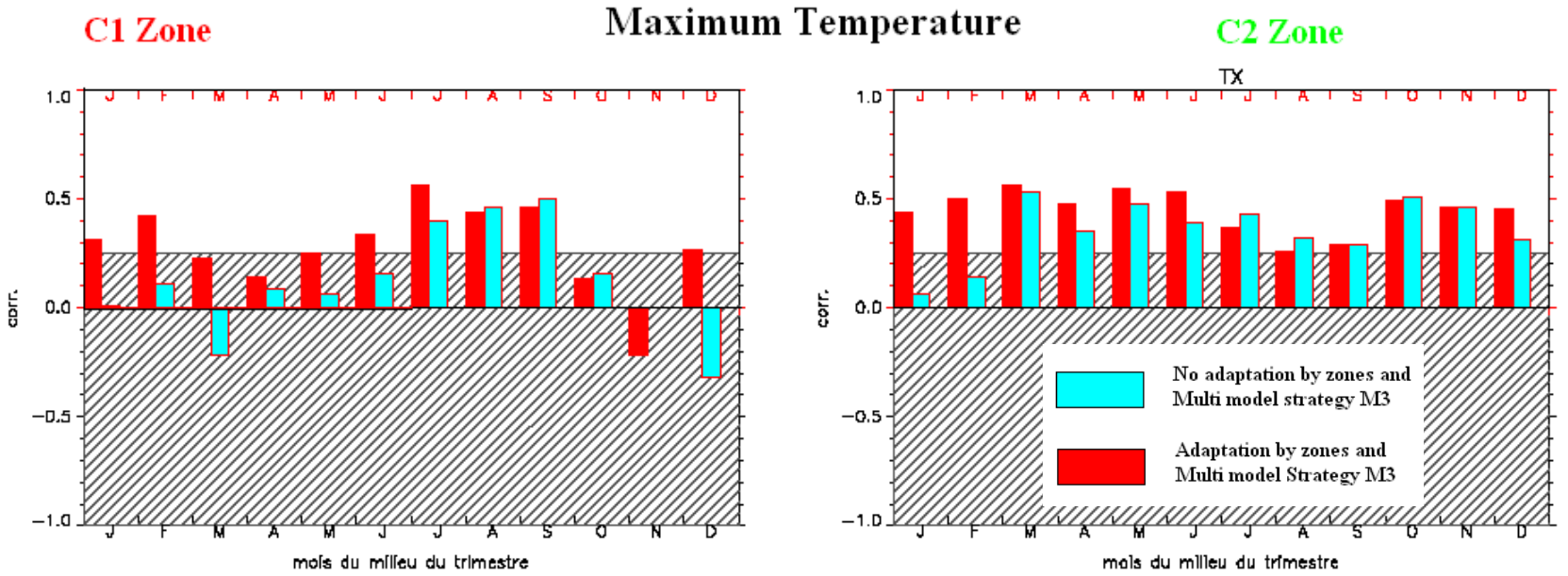
a) **Rainfall**

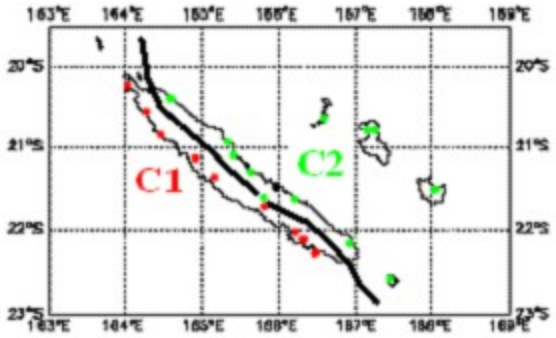
b) **Minimum Temperature**

c) **Maximum Temperature**

Downscaling

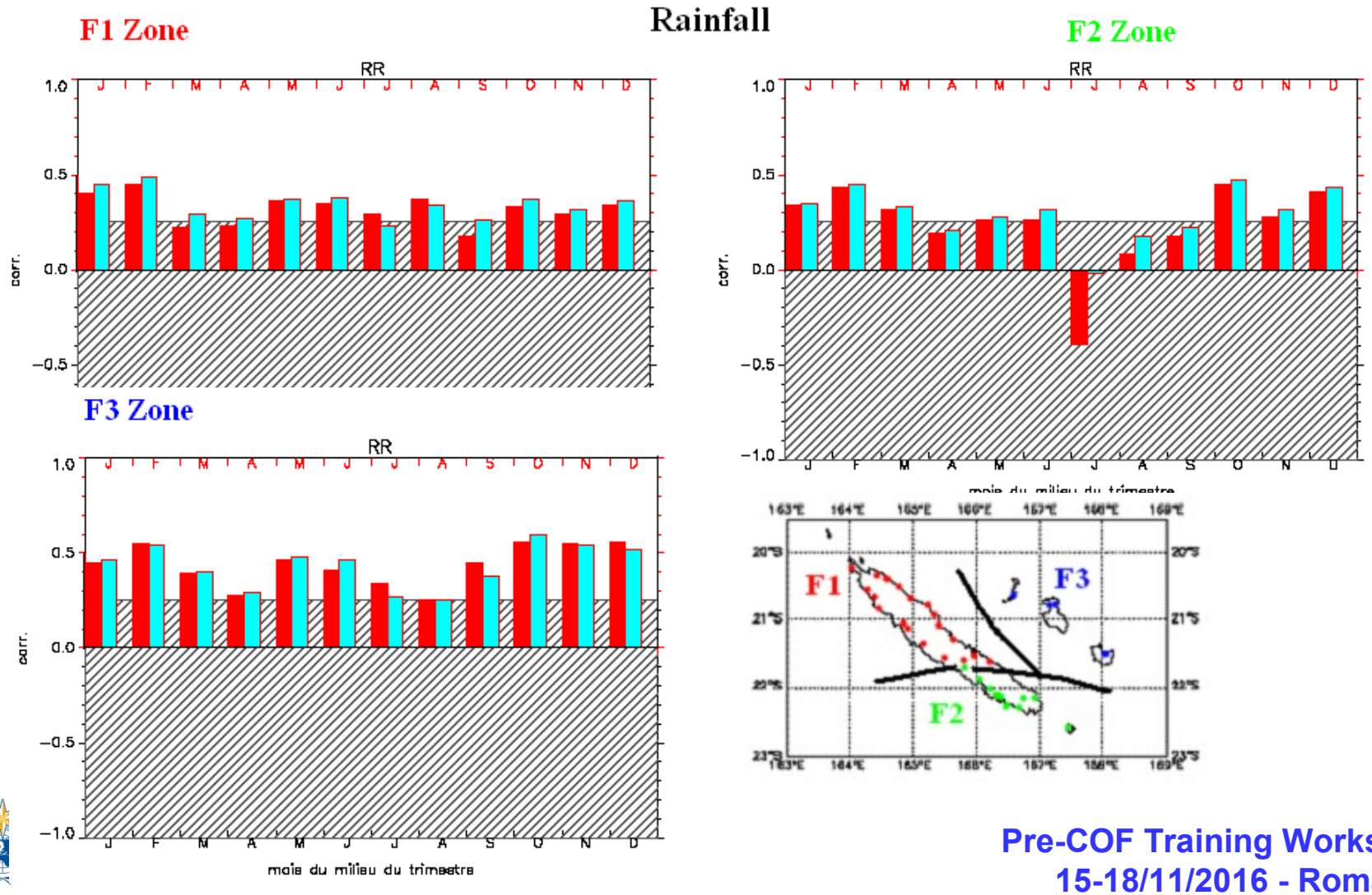
- *Additional information brought by downscaling*





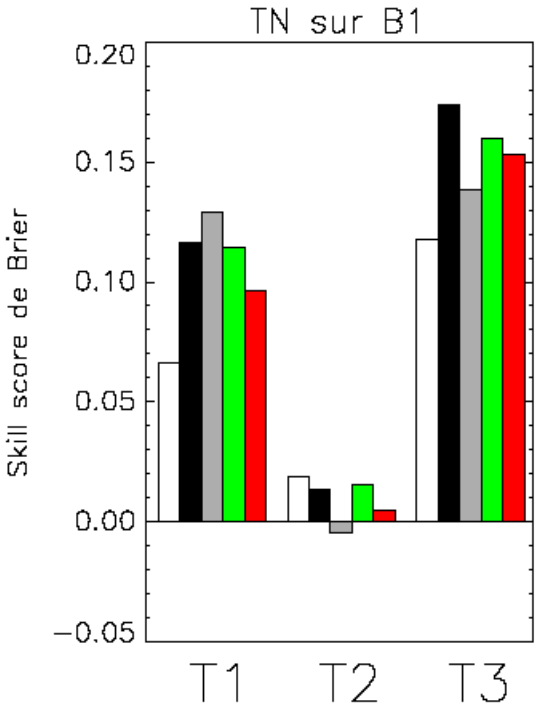
Downscaling

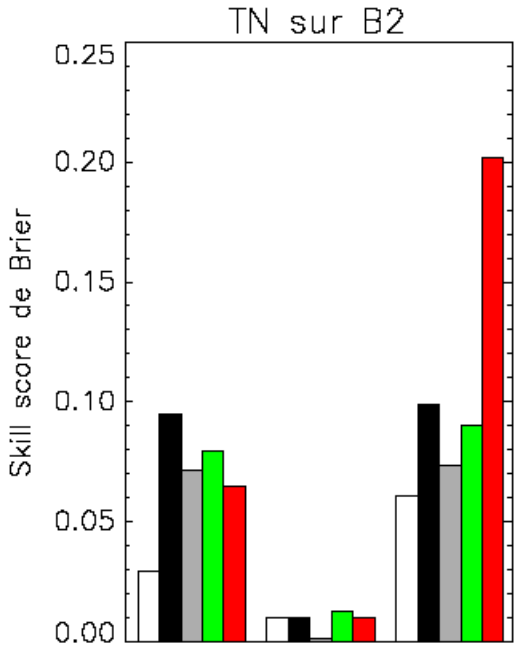
- *Additional information brought by downscaling*

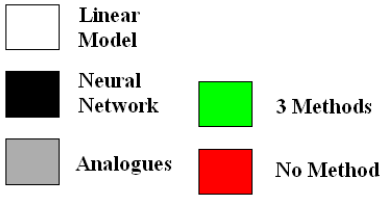


Downscaling

- *Model Choice*
- *Robustness concerns*



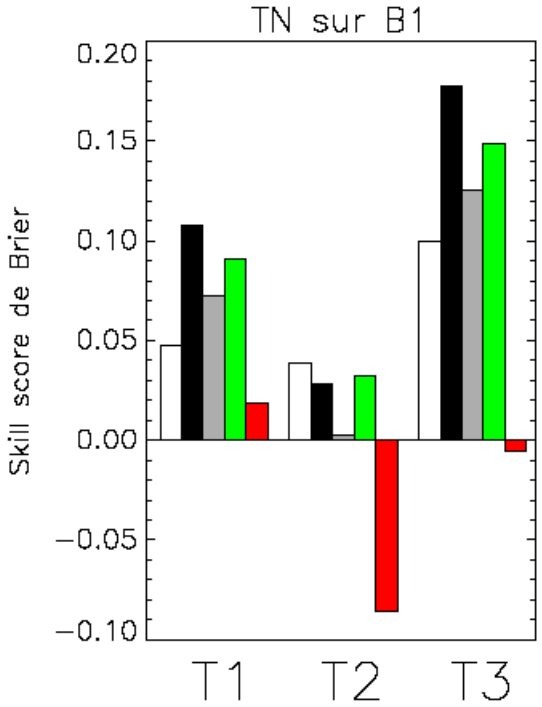


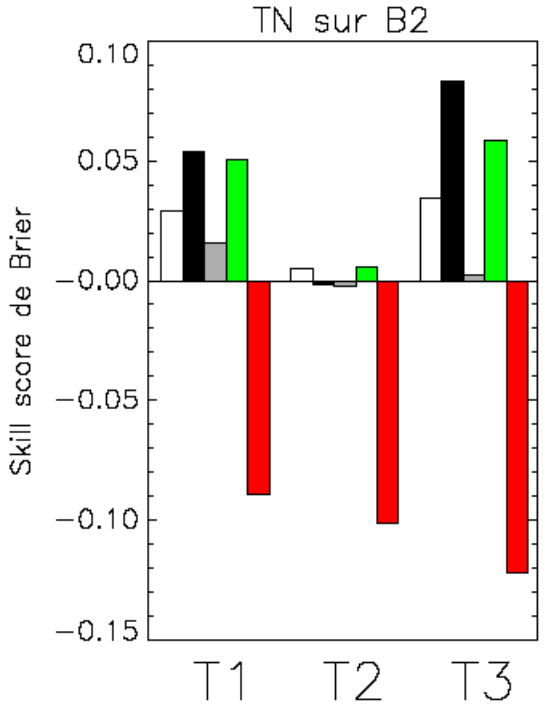


Test on 1980-2002

Minimum Temperature

Test on 1958-1979



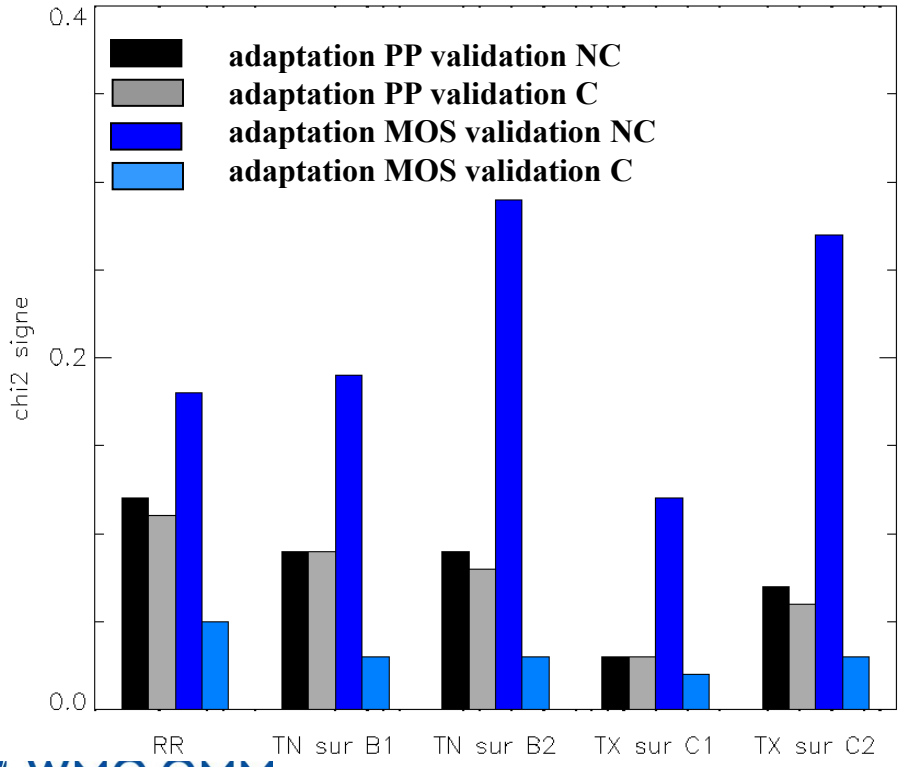


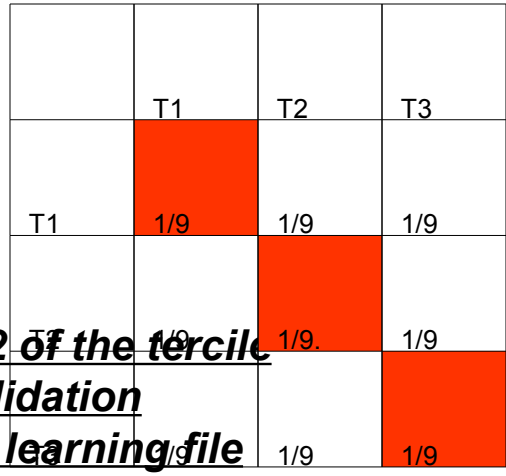
WMO OMM

Downscaling

- Comparison of MOS adaptation using a 15-year climatology with a Perfect Prog adaptation using reanalysis (or longer hindcasts ?)

« Arithmetic Chi2 »





Arithmetic Chi2 of the tercile forecasts in validation modes : NC on learning file C cross vaildation
For Perfect Prog mode (PP) and MOS mode (MOS) over a 15-year period.

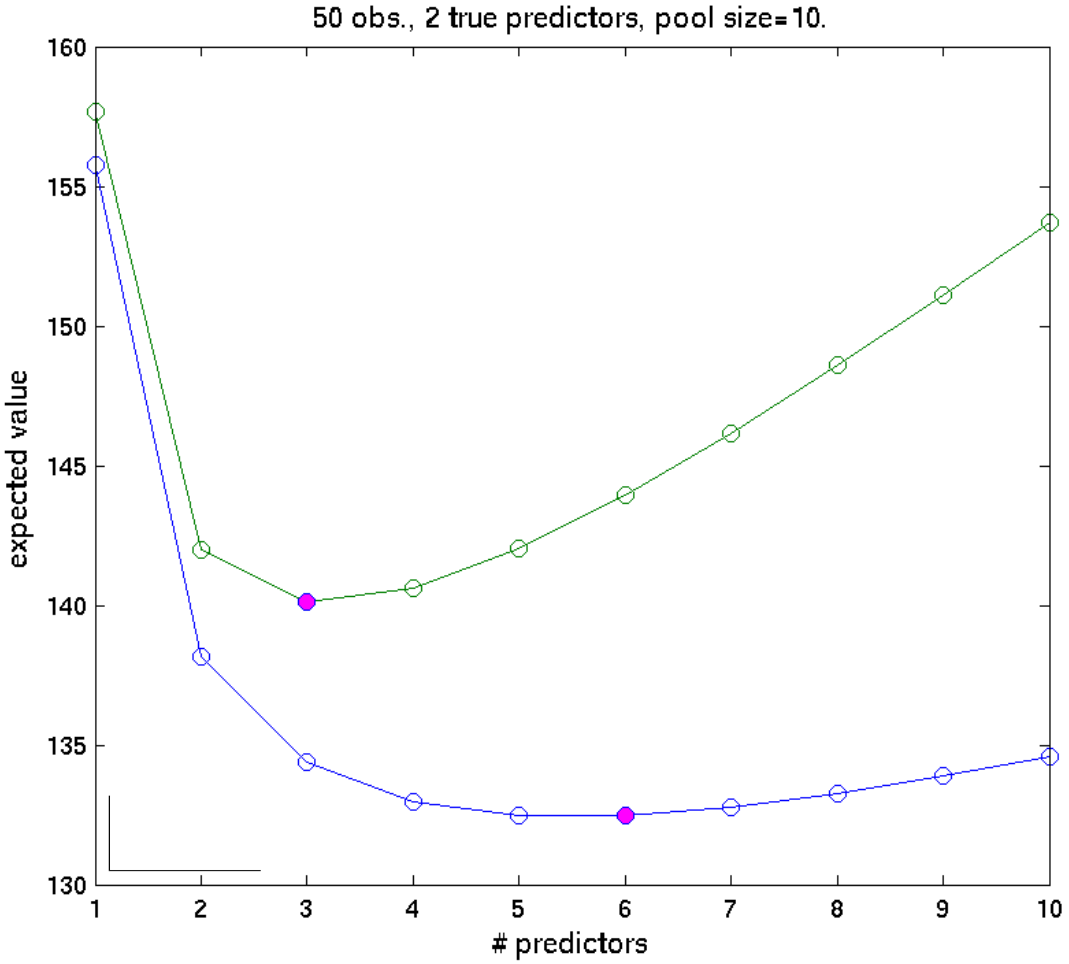
Cross-Validation

Cross-validation is not very good at determining whether the skill score is reliably estimated, but leave-k-out cross-validation can be used to determine whether the right model has been selected (as long as there are not too many candidate predictors). So why not use cross-validation as a selection procedure rather than verification procedure?

Cross-validation as a model selection procedure involves selecting the model with the best predictive capability rather than the best descriptive capability.

Cross-Validation

Model selection criteria indicate improved model selection for leave-k-out over leave-1-out cross-validation.



WEATHER CLIMATE WATER
TEMPS CLIMAT EAU

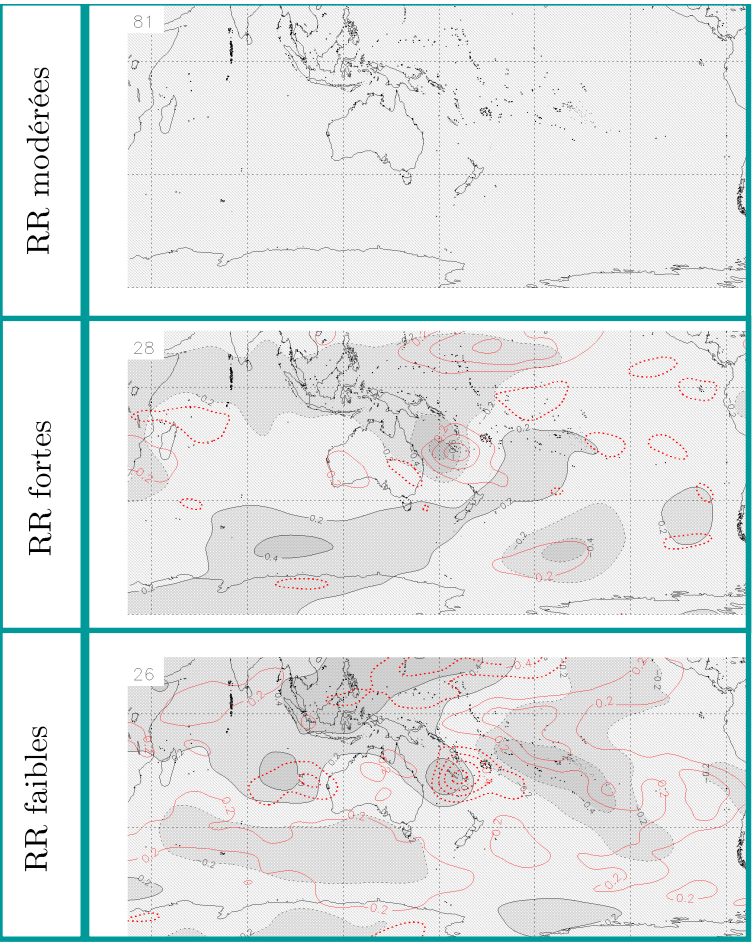
WMO OMM

World Meteorological Organization
Organisation météorologique mondiale

MOS Predictors

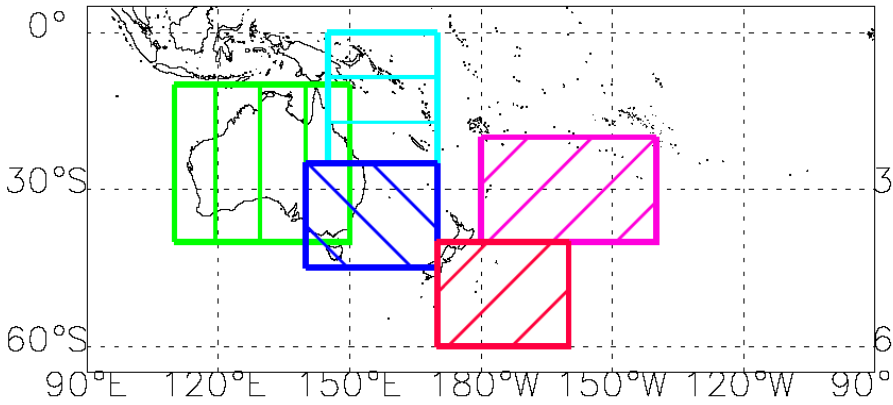
- **Identification of Sources of Large Scale Forcing**

Standardized Anomalies of SLP and Hu700 in JFM vs local anomalies



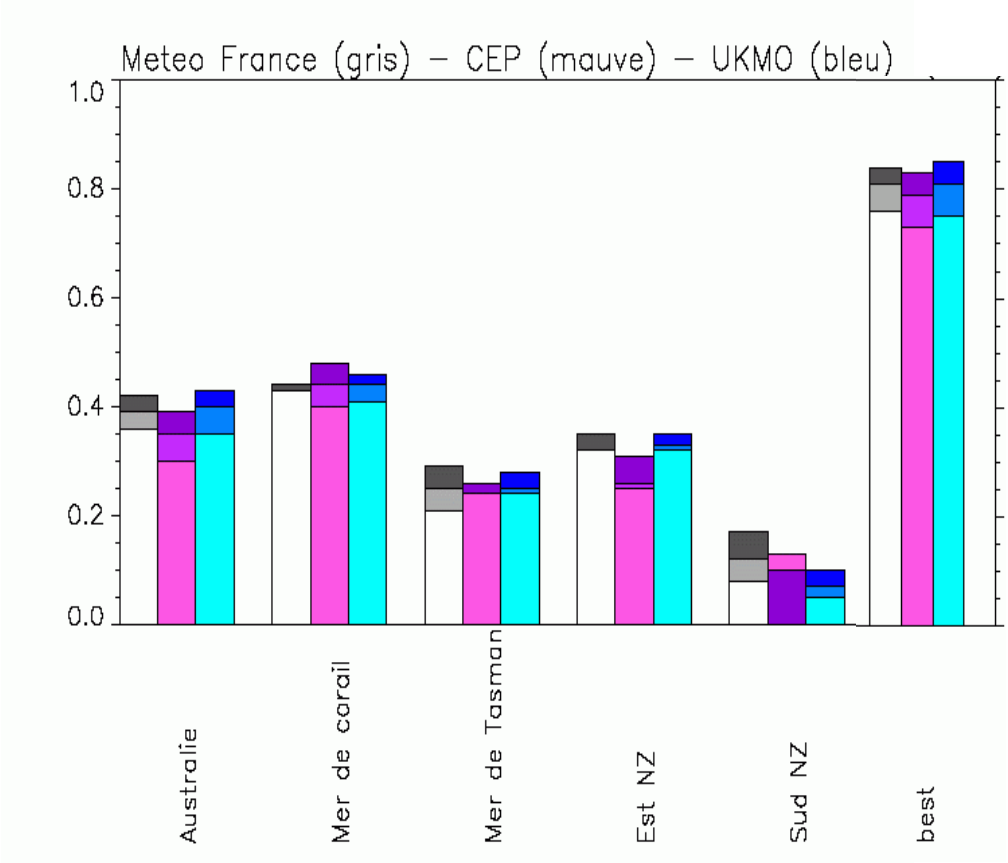
- *Method : composite analysis of large scale fields by anomaly categories of « local » parameters (here in neutral conditions over the Pacific)*
- *Résults : identification of key regions*

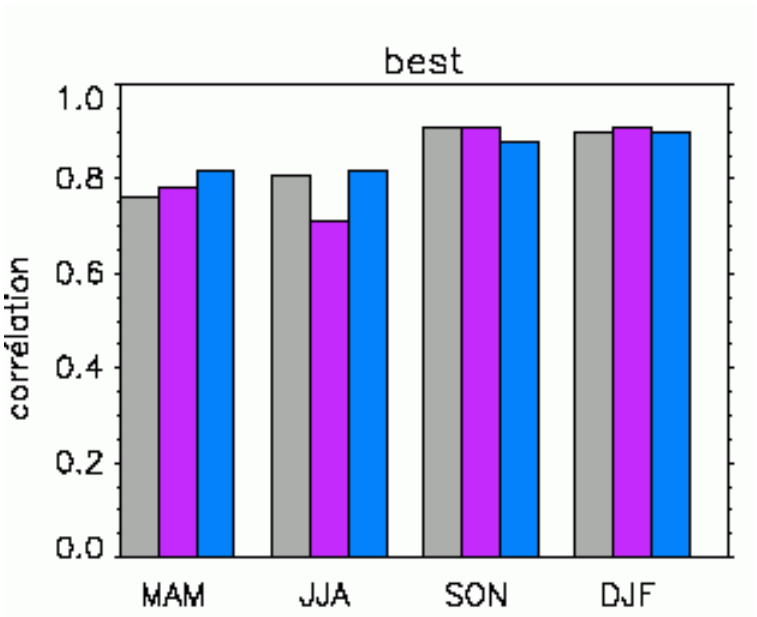
5 zones where the signal of SLP is significant



MOS Predictors

- *Are the predictors well forecasted by the GCM ?*





Corrélation between forecasted indice and reference