Consolidation methods for monthly forecasts from Multi-Model Ensembles (MME) Part 1

Malaquias Peña and Huug van den Dool

Consolidation

- Making the best single forecast out of a number of forecast inputs
- Necessary as large supply of forecasts available
- Expressed as a linear combination of participant models:

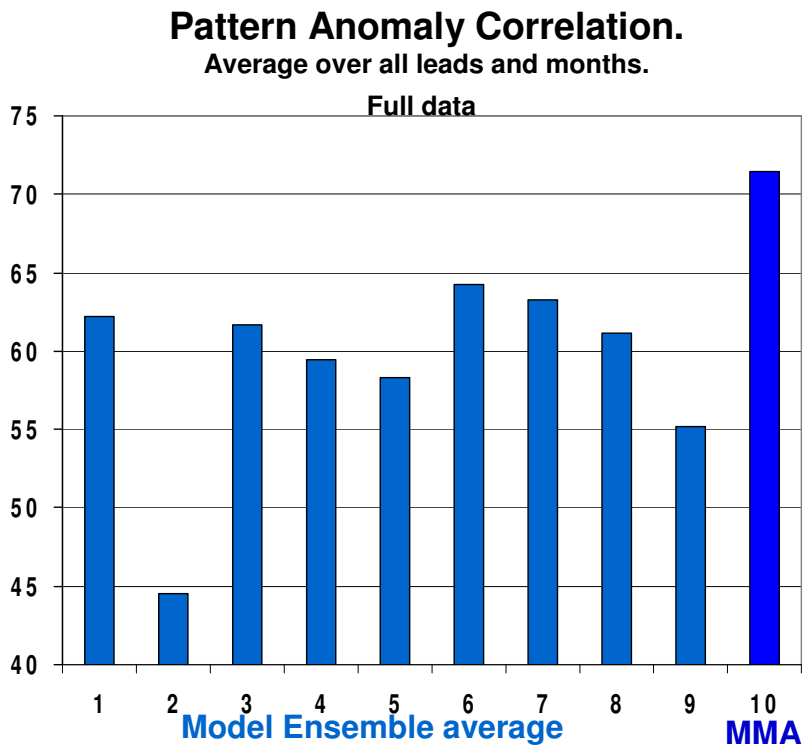
$$C = \sum_{k=1}^{K} \alpha_k \zeta_k$$

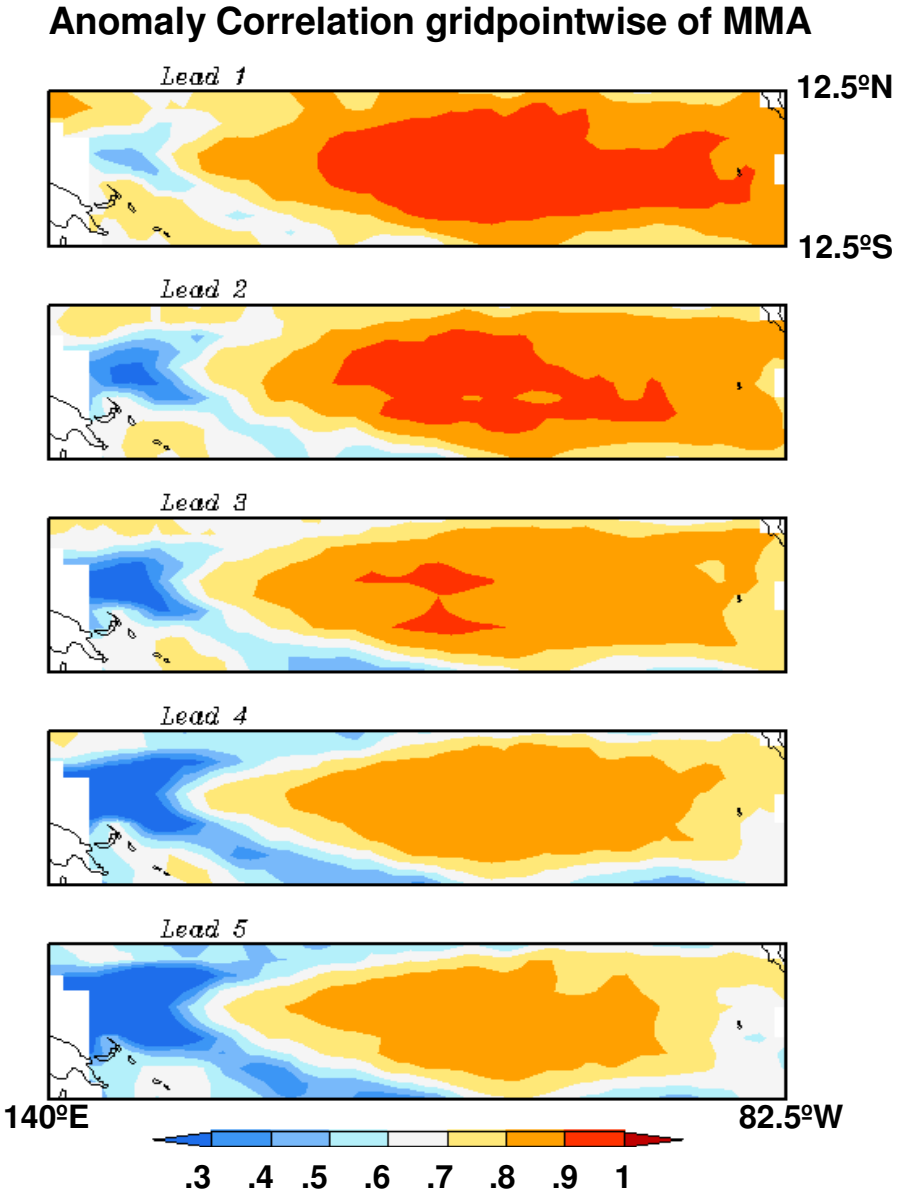
- K number of participating models
- ζ input forecast at a particular initial month and lead time

Task: Finding *K* optimal weights, α_k , corresponding to each input model

- **Data:** *Nine ensemble prediction systems* $(DEMETER+CFS+CA)$
  - At least 9 ensemble members per model
  - Hindcast length: Twenty-one years (1981-2001)
  - Monthly mean forecasts; Leads 0 to 5 months
  - Four initial month: Feb, May, Aug, Nov

Tropical Pacific SST



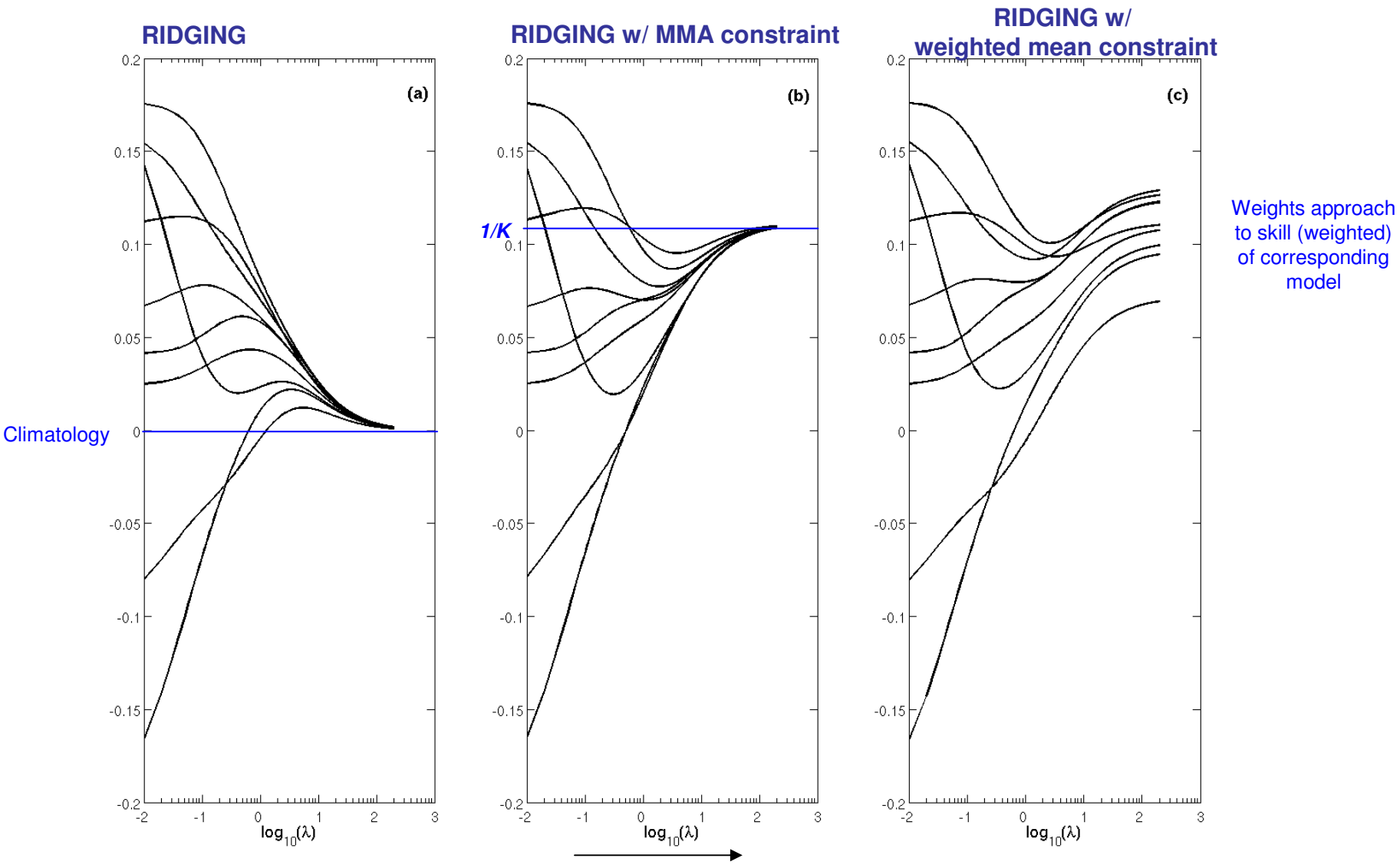


- Region of appreciable skill
- Multi-Model Ensemble Average (MMA) more skillful than any single ensemble model average
- Can sophisticated consolidation methods be better?

Issues

- **HINDCAST should be sufficiently long to**
  - Remove Systematic Errors
  - Carry out weight optimization procedures (consolidation)
  - Allow cross-validation assessment
    - A CV-1 produce degeneracy; a CV-3yrs out will leave only 18 data points.
- **OVERFITTING:** Due to short training dataset compared to the number of input forecast models.
  - Rigorous cross-validation procedure to measure realistic skill
- **COLLINEARITY:** A large number of participating models may lead to problems when at least one of the models is not independent from the rest.
  - Even with plentiful data
  - Covariance matrix ill-conditioned: Regression coefficients not accurately computed
  - Regularization methods required

Ridging Methods



Optimized weights for the 9 model ensemble averages as a function of the ridging amount: λ

Consolidation Methods Assessed

Acronym	Method	Characteristics
MMA	Multi-method ensemble average	Simple average of all the input forecasts
UR	Unconstrained regression	Typical multiple linear regression. Assumes no collinearity among models. Weights sometimes negative and too large
COR	Correlation coefficient	Skill weighted method. Collinearity among models is not considered
FRE	Frequency of the best	Gives weights depending on how many times the model has been the best in the training period.
RID	Ridging	Considers both skill and collinearity among models
RI2	Double pass Ridging	First pass to identify unskillful and/or redundant models. Then, set their corresponding weights to zero and perform a second pass with the reduced set of models.
RIM	Ridging with MMA penalty	Ridging with a penalty term for weights departing from MMA
RIW	Ridging with COR penalty	Ridging with a penalty term for weights departing from COR

OPTIMIZING WEIGHTS

- Find weights, α_i ,for each forecasting tool, ζ_i, that minimizes the (sum of square of) errors ε_j in

$$\mathbf{Z}\vec{\alpha} = \vec{o} + \vec{\varepsilon}$$

Where $\mathbf{Z}$ is a matrix whose columns are the forecasting tools and rows are the data points in the training period, $\boldsymbol{o}$ is the column vector containing the verifying field, and $\boldsymbol{\varepsilon}$ is a vector of errors.

- Least square method (unconstrained regression):

$$SSE = (\mathbf{Z}\vec{\alpha} - \vec{o})^T (\mathbf{Z}\vec{\alpha} - \vec{o})$$

$$\vec{\alpha}_{UR} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\vec{o}$$

ILL-POSED MATRIX PROBLEM

$$\vec{\alpha}_{UR} = (Z^T Z)^{-1} Z^T \vec{o}$$

$(Z^T Z)^{-1}$ eigenvalues	Nino 3.4	PNA	NAO
1	8.4584	5.9156	3.6889
2	0.1763	0.8394	1.402
3	0.1516	0.7808	1.1173
4	0.0707	0.42	0.8759
5	0.0536	0.3488	0.6316
6	0.0384	0.2874	0.5277
7	0.0297	0.1919	0.3978
8	0.0186	0.139	0.2462
9	0.0027	0.0772	0.1126

$\sum \alpha_i^2$ too large

Corresponding weights for UR for lead 1, im 1

1	2	3	4	5	6	7	8	9
0.482	0.2532	-0.5526	-0.5615	0.0189	0.0348	0.018	0.0381	0.0488

RIDGE REGRESSION

Minimize: $SSE = (Z\alpha - o)^T (Z\alpha - o)$

Constrained to: $\alpha^T \alpha < c$ leads to

$$\vec{\alpha}_{RID} = (Z^T Z + \lambda I)^{-1} Z^T \vec{o}$$ Ridge Regression

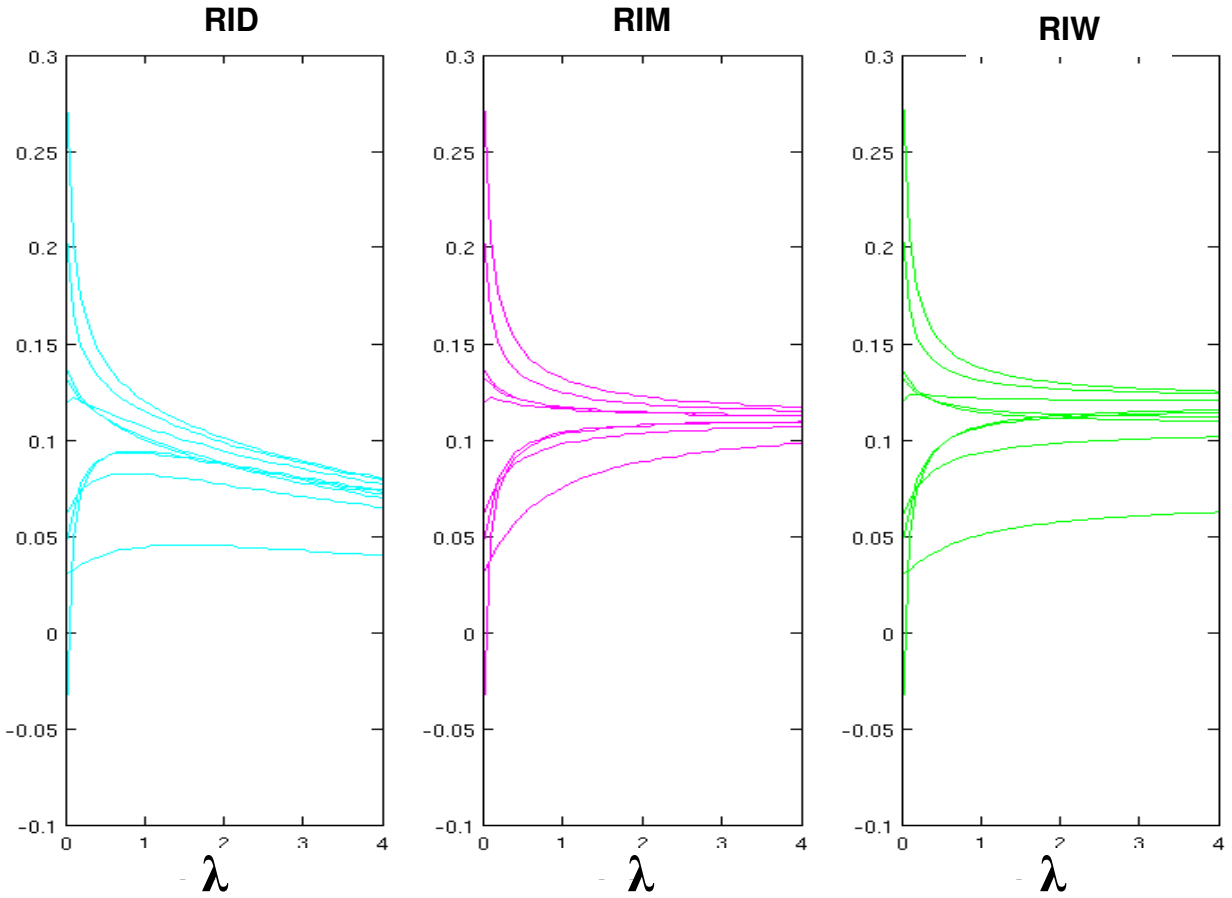
$$\vec{\alpha}_{RIM} = (Z^T Z + \lambda I)^{-1} \left(Z^T \vec{o} + \frac{\lambda}{K} \vec{1} \right)$$ (DelSole, 2007)

$$\vec{\alpha}_{RIW} = (Z^T Z + \lambda I)^{-1} \vec{b}^*$$ (ad hoc)

where $\vec{b}^* = o_i \zeta_i \left(1 + \frac{\lambda}{\zeta_i^2 f} \right)$ and $f = \sum_{i=1}^{K} \frac{o_i \zeta_i}{\zeta_i^2}$

- Van den Dool estimates λ such that the weights are small and stable
- Many more ways to find it
- Depends on characteristics of covariance matrix $Z^T Z$

RIDGE REGRESSION



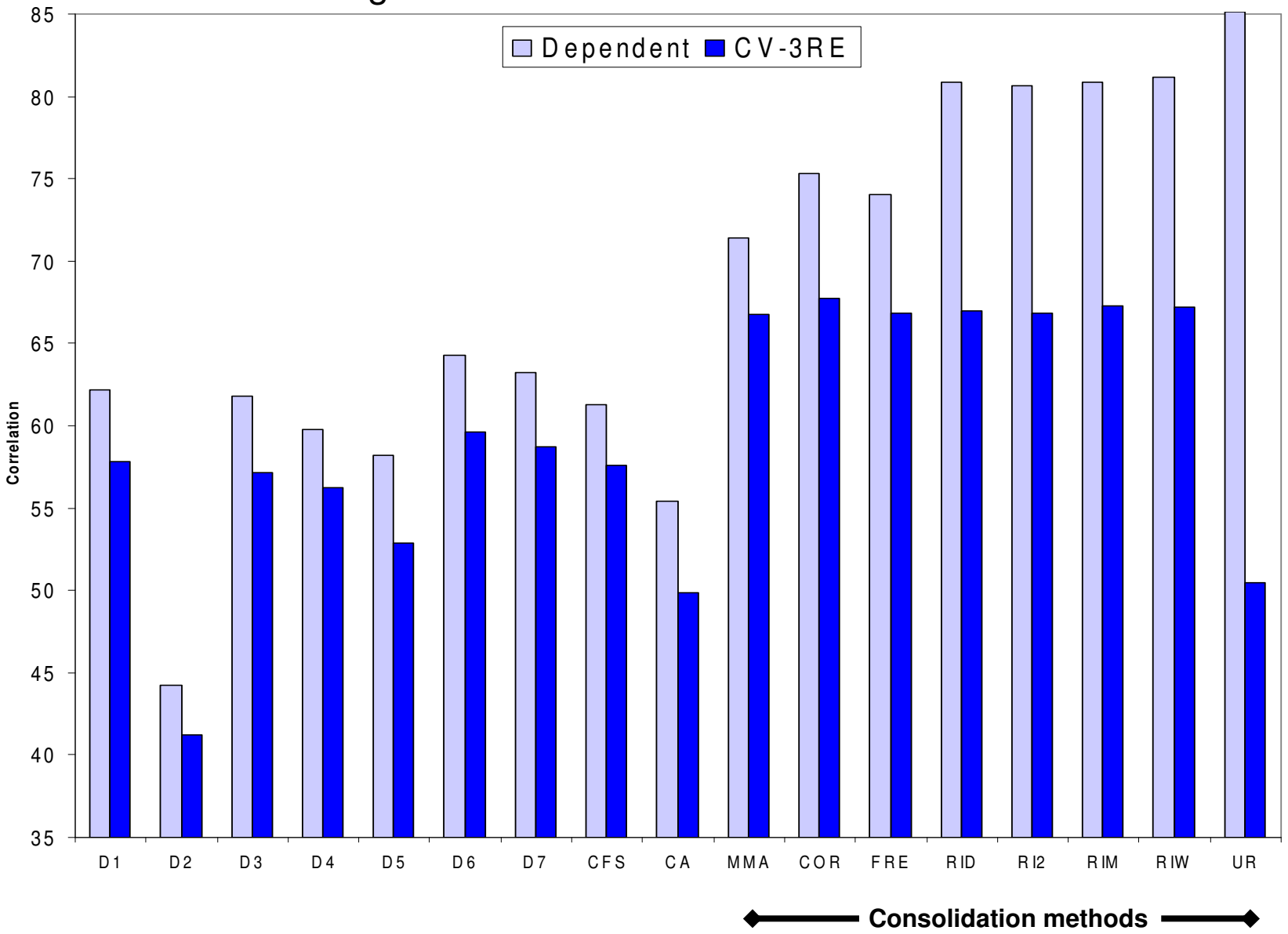
- Model weights (α_i, $i=1..9$) as a function of λ for three ridge consolidation methods.
- Figure illustrates asymptotic values. Our methods stop at $\lambda=0.5$.
- Unconstrained regression ($\lambda=0$) results in a wide range (including negative values) of weights.

CONSOLIDATION METHODS ASSESED

Multi-model ensemble mean (MM)	$\alpha_i = 1/9$, $i=1,..K$, K number of methods
Correlation (COR)	$\alpha_i = \frac{\mathrm{cov}(\zeta_i, O)}{\sigma^2_{\zeta_i}}$, ζ_i time series forecast of *i-th* method
Frequency of best (FRE)	$\alpha_i = \{N_i / N\}$, N number of training years, $N_i = \left\{ \sum_N cases(\zeta_i) \,\middle|\, \zeta_i = \min\{(\zeta_k - O)^2, k = 1,..K\} \right\}$
Ridging (RID)	$\vec{\alpha} = (\mathbf{A} + \lambda \mathbf{I})^{-1}\mathbf{b}$ $\mathbf{A} = \mathbf{Z}^T\mathbf{Z}$, $\mathbf{b}=\mathbf{Z}^T\mathbf{O}$, λ is such that $\alpha_i \geq -0.01$, $i=1,..,K$ *and sum alpha squared small*
Double pass Ridging (RI2)	Set to zero any $\alpha_i < 0$, $i=1,...,K$ *after first RID pass.*
RID with MM constraint (RIM)	$\vec{\alpha} = (\mathbf{A} + \lambda \mathbf{I})^{-1}\left(\mathbf{b} + \frac{\lambda}{\mathrm{K}}\mathbf{1}\right)$
RID with weighted mean constraint (RIW)	$\vec{\alpha} = (\mathbf{A} + \lambda \mathbf{I})^{-1}\mathbf{b}^*$ where $b_i^* = b_i\left(1 + \frac{\lambda}{a_{ii} f}\right)$ and $f = \sum_{i=1}^{K} \frac{b_i}{a_{ii}}$.
Unconstrained (UR)	$\vec{\alpha} = \mathbf{A}^{-1}\mathbf{b}$

Pattern Anomaly Correlation

Average over all leads and initial months



INCREASING EFFECTIVE SAMPLE SIZE

APPROACHES

1. Selection-combination (double pass strategy)
   - Objective procedure to remove or set to zero weights of bad or redundant models
   - First pass: Ridging identifies negative weights. Set these to zero
   - Second pass: Ridging is carried out on the models with positive weight
2. Mixing data from neighboring gridpoints
   - Previous studies show a gain in weight stability
   - Reduces flexibility in cases where the raking of the model skill changes from region to region
   - If hindcasts allowed, mixing neighboring lags
3. Mixing information from individual ensemble members
   - Most studies have used the ensemble average for each model
   - Each member is a unique realization of model
   - Constrain weights of ensemble members to be same within each model

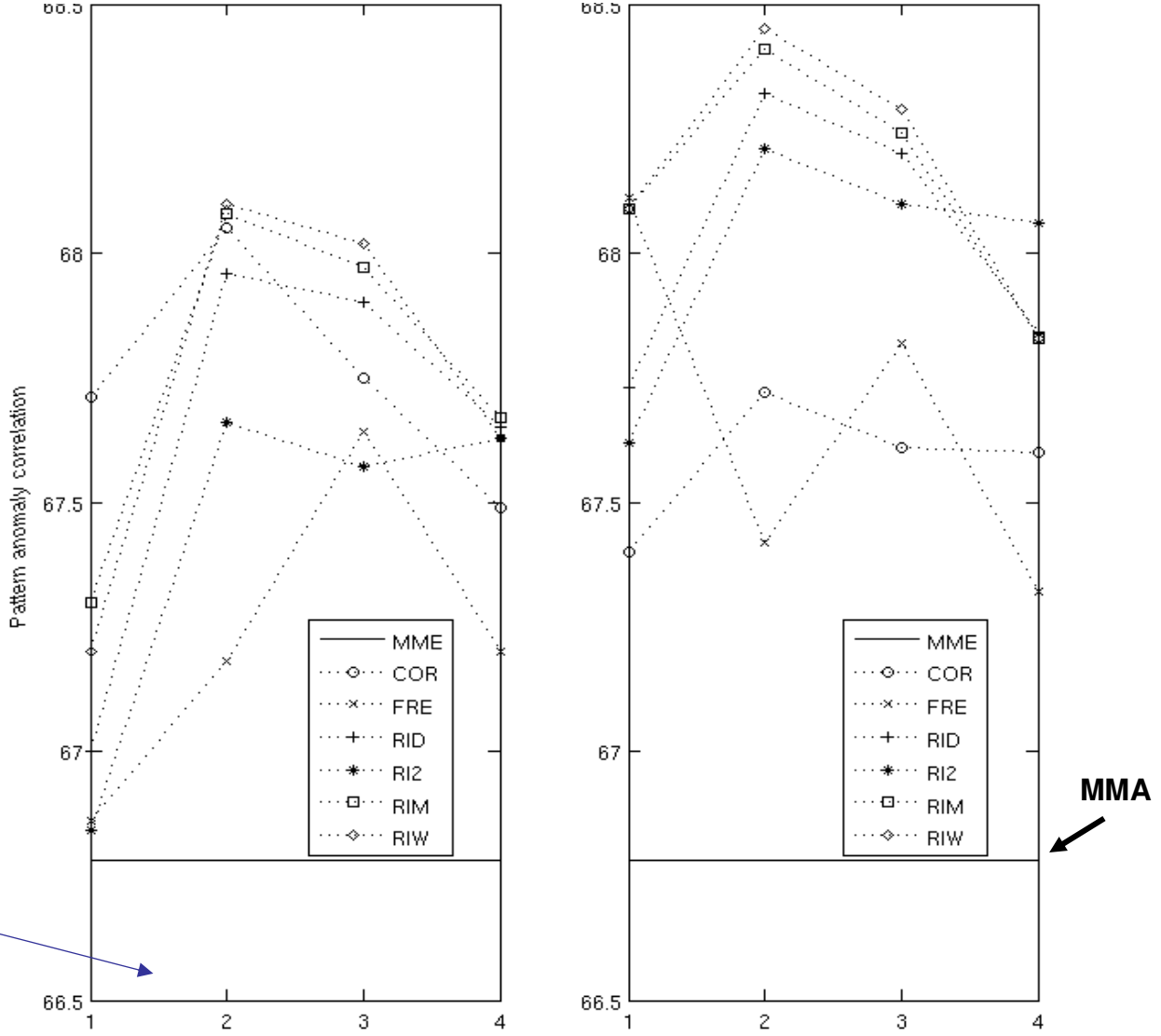
INCREASING EFFECTIVE SAMPLE SIZE

Approaches

1. Gridpoint by gridpoint
2. 3x3 box that includes the point of analysis plus the 8 closest gridpoints
3. 9x9 box with the 80 closest neighboring gridpoints
4. All the grid points in the domain

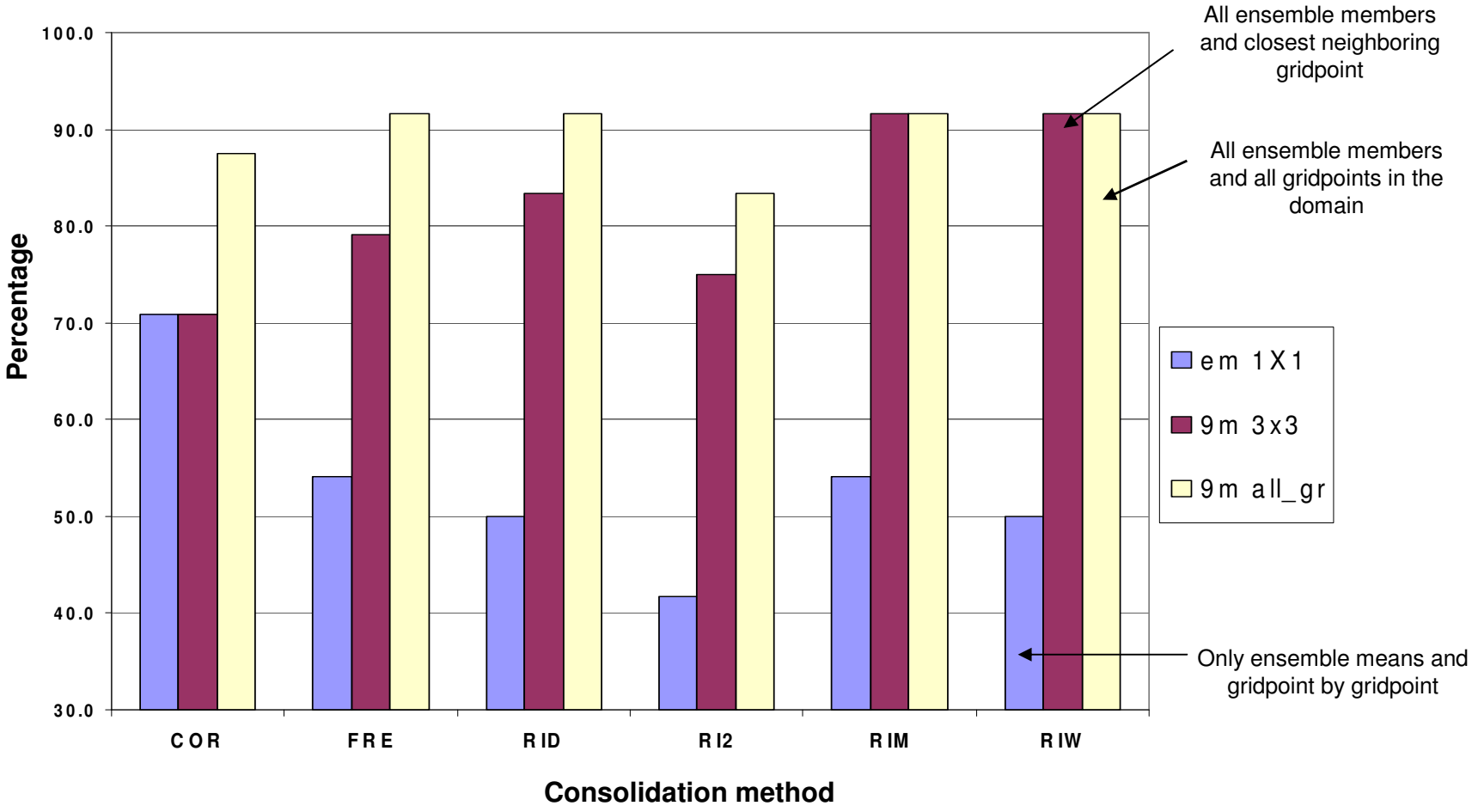
USING ENSEMBLE AVERAGES ONLY

USING ALL ENSEMBLE MEMBERS



INCREASING EFFECTIVE SAMPLE

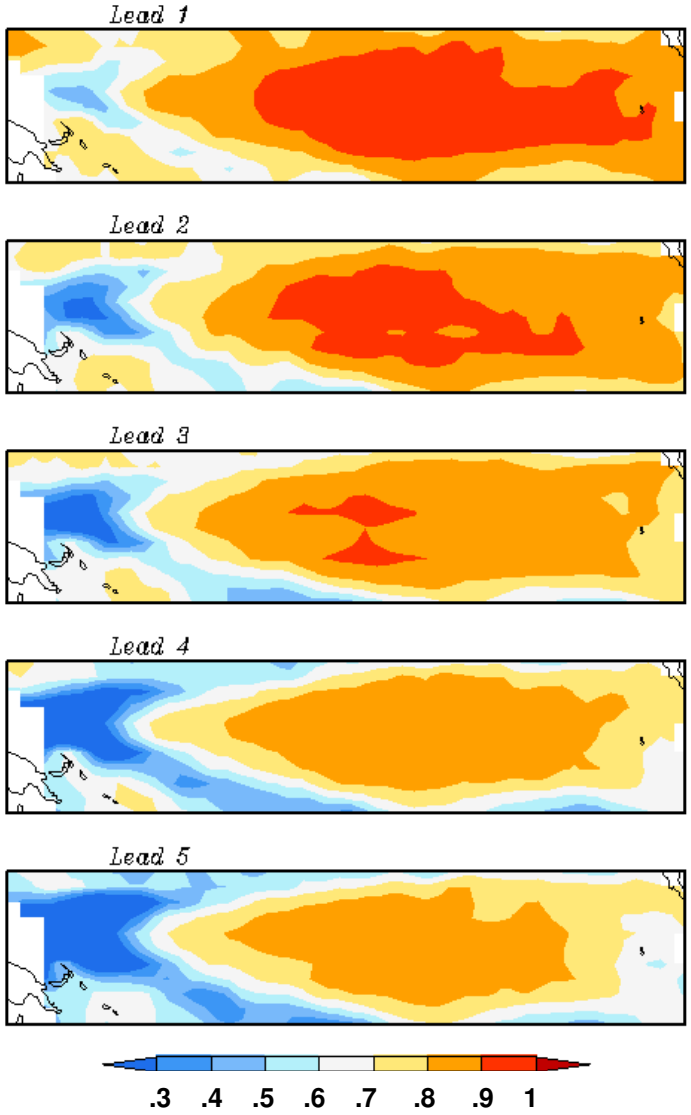
Consistency: Percentage number of cases that a sophisticated consolidation method outperforms MM



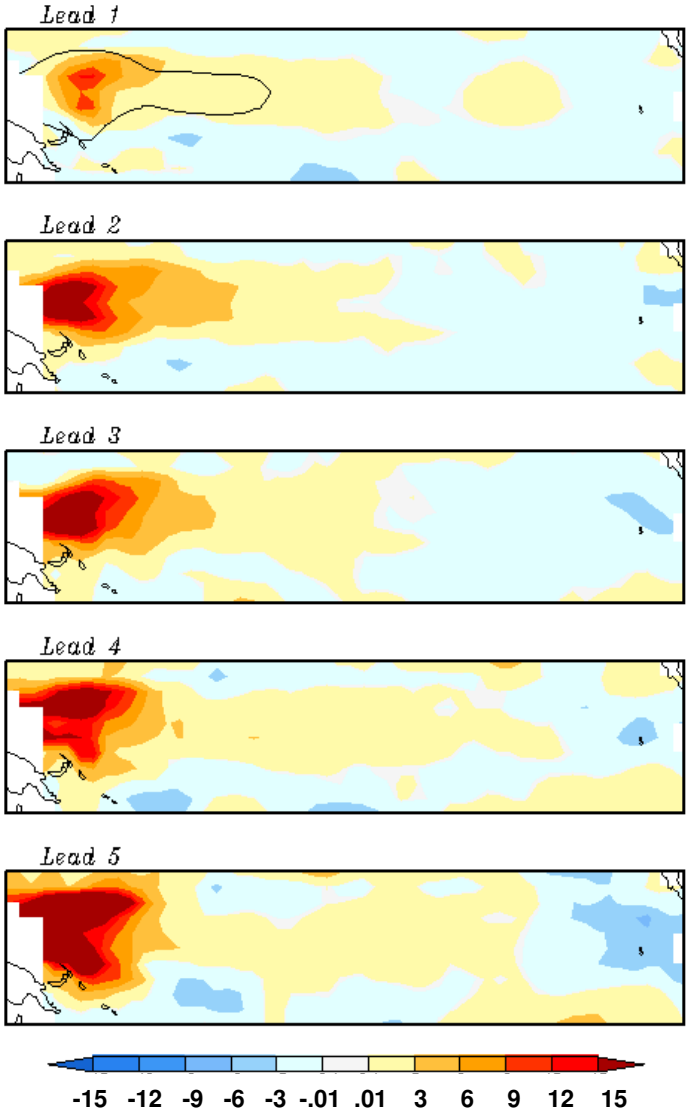
Sophisticated consolidation methods more frequently outperform MMA for larger effective sample

Performance

AC for MMA. All initial months included



AC: RID minus MMA (%). All initial months included



Ridging AC improves considerably over western Pacific

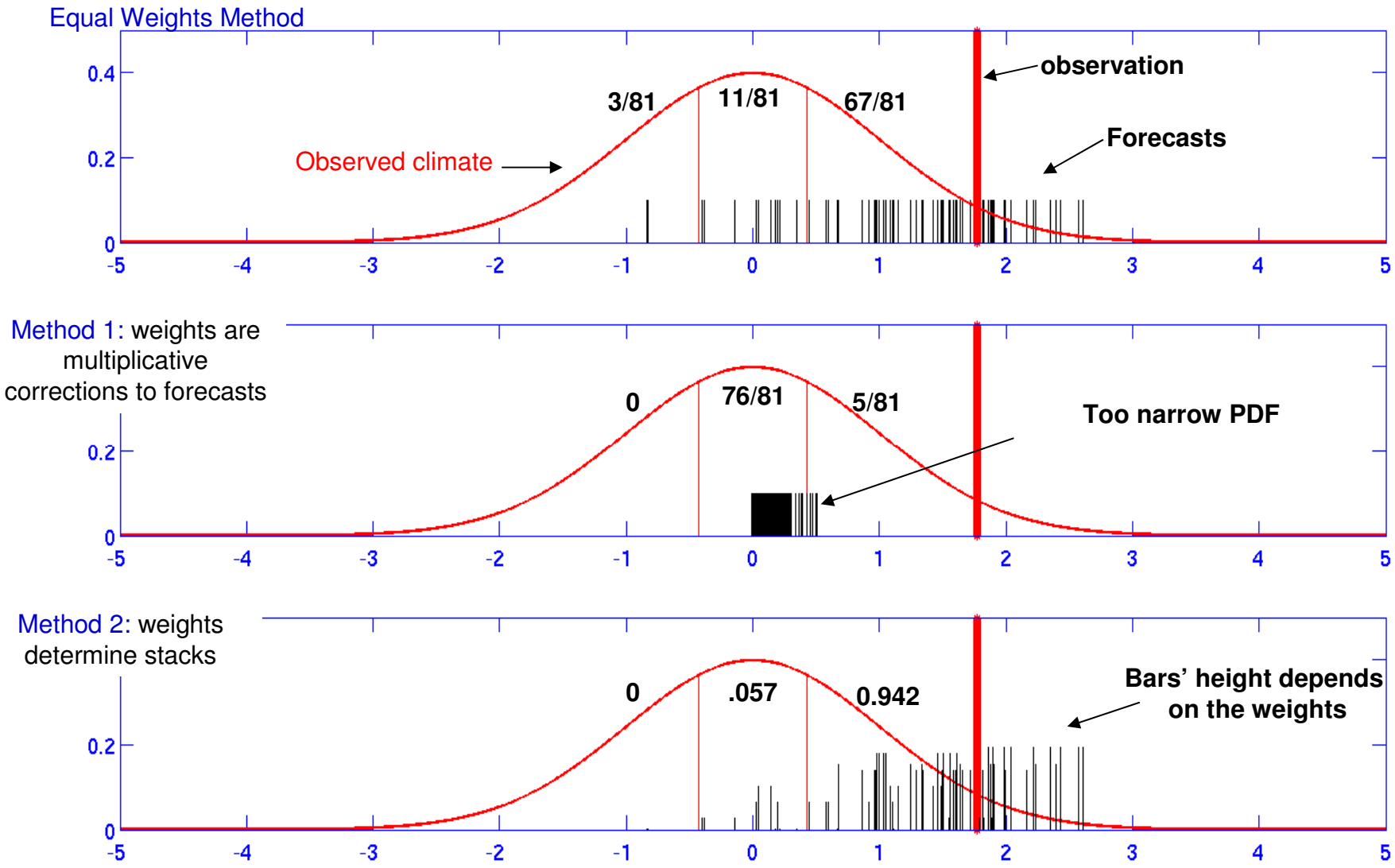
Probabilistic assessment

- PDF formation from optimized weights
    - Method1: Weights used as factors that multiply corresponding forecast values.
        - Width of PDF becomes too narrow when weights are small.
        - Optimization to inflate PDF required
    - Method2: Weights used to determine stacks given to corresponding model without changing the value of the forecast
    - Both methods produce similar results with the latter more strightforward
- Relative Operating Characteristic Curve:
    - To measure the ability of method to predict occurrence or non-occurrence that observation will fall in the “upper”, “middle” or “lower” tercile.
    - Class limits defined by the observed SST during the training period

PDF formation

Illustration for a particular gridpoint, lead and initial month

p = fraction of ensemble members

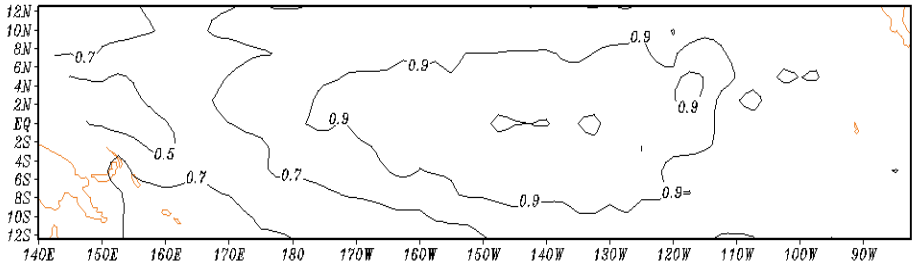


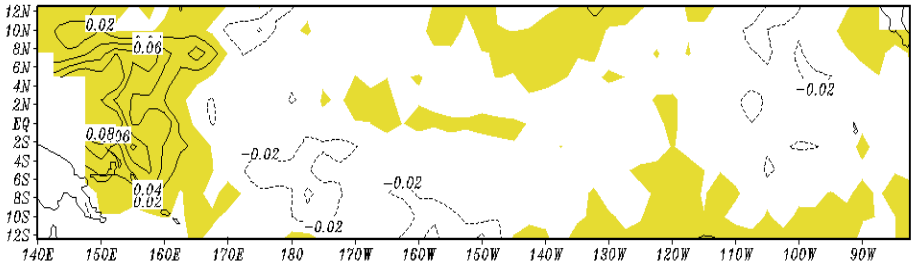
Area below ROC curve

MMA

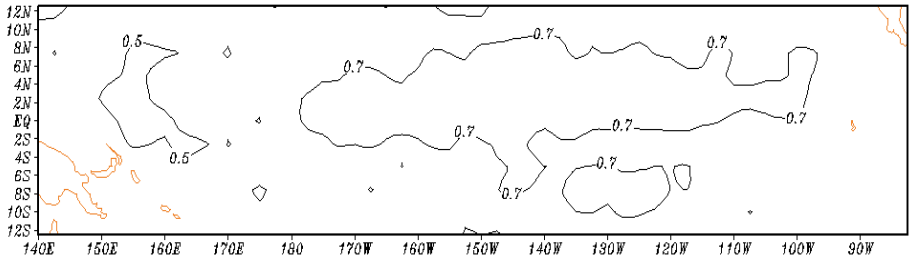
Upper tercile

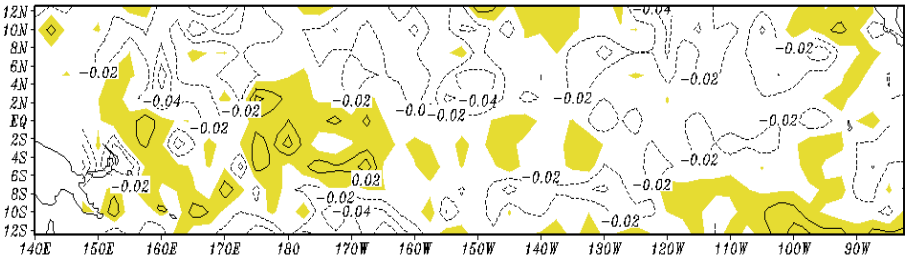
RID minus MMA



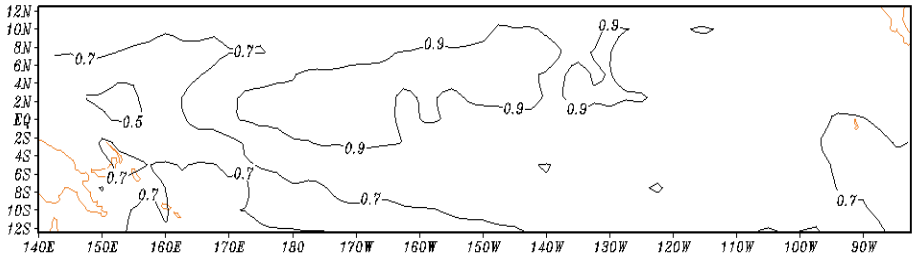


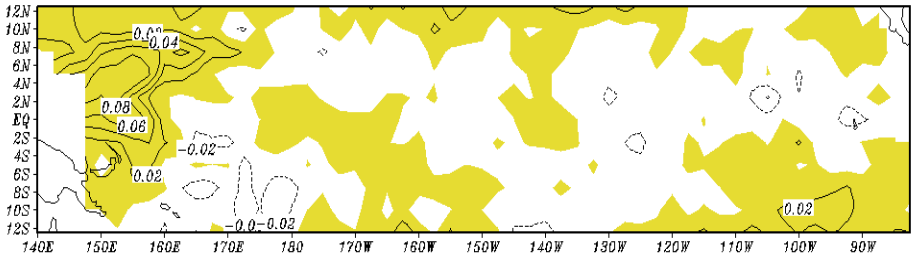
Middle tercile





Lower tercile





Regions where RID outperforms MMA are shaded

Summary of results

- Eight consolidation methods assessed under rigorous cross-validation procedure, CV-3RE, to combine up to 81 monthly forecasts of tropical Pacific SST.
- The simple multi-methods ensemble average (MMA) shows large and consistent skill improvement over individual participant models as measured by AC.
- When the sampling size is small sophisticated consolidation methods are as skillful as (or, in the case of UR, worse than) MMA
- Increasing the effective sampling size produces more stable weights and affects positively the skill of sophisticated consolidation
- In the western tropical Pacific, sophisticated consolidation methods improve significantly over MMA. This is true both using a deterministic (AC) and Probabilistic (ROC) assessment.

Remarks

- Hindcast
  - MMA requires long hindcast to remove SE
  - Sophisticated consolidation methods need sufficiently long hindcast not only to remove SE but to optimize weights
- Cross-validation
  - Rigorous CV-3RE shows artificial skill prevalent in most of the consolidation methods
- Probability Density Function
  - Efficient treatment is needed to adjust posterior PDF in ridge regression methods.
    - Gaussian Kernel, Bayesian Methods
  - SE of the standard deviation and other higher moments of the PDF: hard to obtain given shortness of the hindcasts
- Application dependent
  - Results for SST monthly predictions (high skill and large collinearity) may not apply to consolidation of other variables

Current lines of development

- National MME
  - Ben Kirtman (CCSM3.0, 3.5 and 4.0) with CFS
  - Lisa Goddard (IRI; post-processing methods)
  - Tim delSole (COLA; post-processing)
- International MME
  - European countries and NCEP
- Bayesian Methods for post-processing
  - Distribution Fitter software