Examination of the performance of the Bishop & Abramowitz (2013) independence coefficients for model selection – an idealized experiment

The method of Bishop and Abramowitz (2013) was designed to assign weights according to independence and performance in a model ensemble. It was not designed to select a subset of an ensemble, the purpose for which it is used in this paper. In particular, the method calculates weights (or coefficients) according to a models independence from the rest of the entire ensemble. This does not guarantee that a subset of a few models with the largest magnitude independence weights will be optimally independent from each other.

1. Idealized experiment

An idealized thought experiment demonstrates this.

Assume you want to select two as independent as possible models out of a 5 member ensemble. For the sake of simplicity assume that all ensemble members have the same quality compared to observations (same error variance). Further assume that models 1 and 2 (group A) are identical, models 3, 4, and 5 (group B) are identical as well, and the models in group A are independent from the models in group B. An optimal choice would obviously be to select one model of group A and one model of group B.

But what would be the result of using a ranking based on the Bishop and Abramowitz (2013) independence weights? It would assign the largest weights to both models in group A (with regard to the entire ensemble, they are the most independent ones). I.e. the identical models 1 and 2 would be top ranked and selected. This is clearly not the desired result.

Let us test this idealized thought experiment below.

2. Testing the experiment

Clearly such a situation with identical models could not occur in a climate model ensemble. It is also worth noting that the use of identical models results in a non-invertible covariance matrix and hence this method cannot be applied and independence weights cannot be calculated.

The question then is: how far from identical do the models need to be for the independence weights to identify a model from group A and a model from group B rather than 2 models from the same group?

Here we create two error time series that fulfil the requirements of the thought experiment and then add some random noise to each model so that the covariance matrix is invertible and independence weights can be calculated. The time series created can be seen in Figure 1. One group has an error of 1 for the first half of the record while the other has an error of 1 for the first and last quarters. The noise is added to each model, at each time step, using a random number from a normal distribution centred at 0 and with a standard deviation of 0.01. This standard deviation represents 1% of the range of model errors.

In this case models within their groups have correlations greater than 0.999, while the correlations with models in the other group are less than 0.004.

Based on these time series the two independence weights with the largest magnitude are indeed from the same group as suggested by the thought experiment.

Using this experiment design we can now increase the level of noise relative to the error signal by

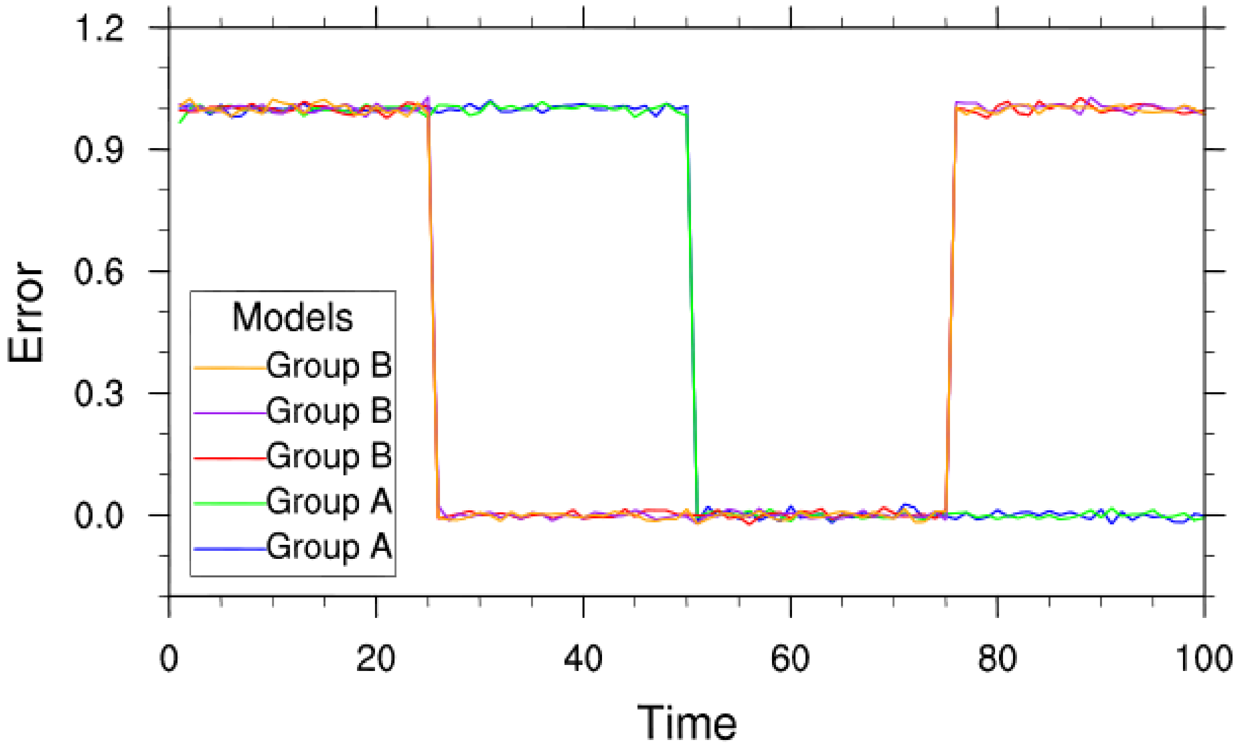

Figure 1: Error time series with random normal noise using a standard deviation of 0.01.

increasing the standard deviation of the normal distribution from which the noise is sampled.

Example time series produced using random normal noise with a standard deviation of 0.1, or 10% of the error signal, is shown in Figure 2.

In this case models within their groups have correlations greater than 0.95, while the correlations with models in the other group are less than 0.04. Example correlation, covariance and inverse covariance matrices are shown below Figure 3, along with the derived coefficients.

Based on these time series the two independence weights with the largest magnitude are now from different groups. Repeating this exercise 1,000,000 times with different random noise reveals that the desired outcome of one model from each group is achieved 87% of the time. This suggests that even with a high signal-to-noise ratio, though some noise is required, in this idealized experiment the independence weights will select optimally independent models almost every time.

A second thought experiment can be considered that requires only that the covariances between models within each group be the same, say 0.9, and the covariances between models in different groups be the same, say 0.1. If the error variances for all models are 1 then these covariances are equivalent to correlations. In this case, as in the first thought experiment, only two unique values for the independence weights will be produced and a sub-optimal model selection will result.

However, if one is dealing with time series that contain some fraction of noise, perhaps from internal variability, obtaining identical covariances are extremely unlikely. The question raised by this thought experiment is “how much variability in covariances, within and between groups, is required to obtain the desired result.” The test example provided here demonstrates that a very small amount of variability in the covariances within and between groups (see Figure 3) results in much larger differences in the inverse covariance matrix and hence the independence weights obtained differ significantly from those obtained from the pure thought experiment. Indeed, in this

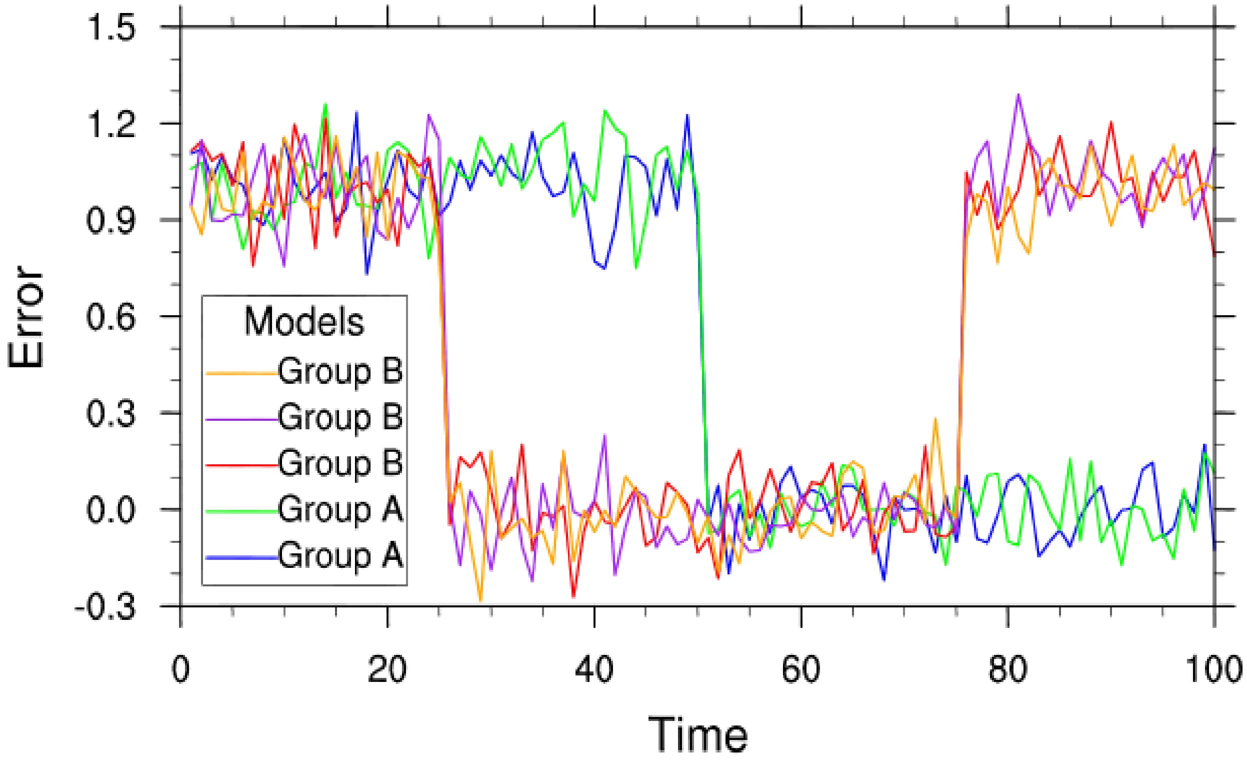

Figure 2: Error time series with random normal noise using a standard deviation of 0.1.

test experiment only a relatively small amount of noise is required for the independence weights to produce the desired model selection most of the time.

Correlation matrix

1	0.9764	-0.0208	-0.0279	-0.0135
0.9764	1	-0.0171	-0.0256	-0.0125
-0.0208	-0.0171	1	0.9754	0.9738
-0.0279	-0.0256	0.9754	1	0.9801
-0.0135	-0.0125	0.9738	0.9801	1

Covariance matrix

0.2669	0.2537	-0.0067	-0.0093	-0.0043
0.2537	0.2598	-0.0054	-0.0084	-0.0041
-0.0067	-0.0054	0.2566	0.255	0.2533
-0.0093	-0.0084	0.255	0.2738	0.2643
-0.0043	-0.0041	0.2533	0.2643	0.2714

Inverse covariance matrix

52.601	-51.3407	2.5037	0.881	-3.1265
-51.3407	53.9796	-2.5993	0.2713	2.1534
2.5037	-2.5993	62.6125	-31.9783	-27.283
0.881	0.2713	-31.9783	77.4949	-45.6044
-3.1265	2.1534	-27.283	-45.6044	73.5317

Independence coefficients

0.1904	0.309	0.4083	0.1335	-0.0412

Figure 3: Example correlation,covariance and inverse covariance matrices from time series generated using random normal noise with a standard deviation of 0.1

3. Implications for use with real climate models

What implications does this idealized thought experiment have for real climate model ensembles? One possible situation where the use of the independence weights to select models will be sub-optimal can be identified using the ensemble correlation matrix. If the models separate into groups such that within each group they are extremely highly correlated, while models in different groups have almost no correlation, then this selection method will be sub-optimal. The levels of correlation required within a group are however extremely high (above 0.96), while those between groups are extremely low (below 0.03). When one considers that these levels need to be achieved across the domain for multiple variables (here daily precipitation, maximum and minimum daily temperature), it is highly unlikely that such a situation could occur with actual climate model data.

The model ensemble considered in this study shows correlations for precipitation ranging from 0.05 to 0.8, while correlations for maximum and minimum temperature range from 0.75 to 0.99. There is also no identifiable group with very high correlations within the group and very low correlations with models outside the group.

While the specific example considered in this idealized thought experiment is very unlikely to occur in a real climate model ensemble, it does serve to emphasise that the independence weights derived by Bishop and Abramowitz (2013) were not designed for this model selection purpose. Other situations, beyond that considered here, may occur that would result in a sub-optimal model selection. When tested on real climate model data this method of model selection has been found to select models which act independently compared to model selection based on performance alone (Evans et al 2013). However, this does not preclude the possibility of sub-optimal model selection on a different, as yet untested, model dataset. This method has the advantages of explicitly considering independence and only requiring the same data as most performance metrics require to calculate.

4. Acknowledgements

We would like to thank A. Gobiet for suggesting the idealized thought experiments explored here.

5. References

Bishop, C. H. and Abramowitz, G.: Climate model dependence and the replicate Earth paradigm, Clim Dyn, 41(3-4), 885–900, doi:10.1007/s00382-012-1610-y, 2013.

Evans, J.P., Ji, F., Abramowitz, G. and Ekstrom, M.: Optimally choosing small ensemble members to produce robust climate simulations, Env. Res. Letters, 8(4), doi:10.1088/1748-9326/8/4/044050, 2013.