

Methods of Multi-Model Consolidation, with Emphasis on the Recommended Three-Year-Out Cross Validation Approach

Huug van den Dool

*Climate Prediction Center, NOAA/NWS/NCEP
 Camp Springs, MD 20746*

1. Introduction

In many contexts with limited data and no patience to wait for new and independent data, one needs to design schemes that mimic the real time forecast situation on a fixed old data set. This is done often nowadays by cross-validation (CV). The purpose of CV is to establish properties of a forecast scheme that would apply on independent future data, for instance to estimate a-priori skill. However, while CV is often a necessity, it may also itself be the source of a problem in evaluating skill. CV is not an exactly defined procedure in general, so let's focus on a situation when systematic error correction is thought to be required. Given N pairs of forecast and verification, say seasonal Nino34 forecasts for 1981-2001, we can set M pairs (M much less than N=21) aside, calculate the systematic error over the N-M cases, then apply the correction to all or some of the M cases left out. This is done exhaustively, so all data is used as (assumed independent) verification at least once. Naturally, researchers want to get away with M=0 or M=1, since it is simpler than M>1, and skill may appear higher that way. Don't we want high skill??? Yes, but not if the assessment misleads us as to the performance in real time.

Dependent data generally overstate the skill level. In this write-up we want to make a strong case for M=3, i.e. keep (at least) three forecast/observed pairs out. This appears to be the right approach in the context of multi-model ensembles, where not only systematic error correction is required but also the determination of weights to be assigned to the participating models.

The procedure we recommend, used in Pena and Van den Dool (2008), is more completely named CV3RE, where CV is cross-validation, 3 means three years left out, R refers to the random choice of two of the three years left out, and E refers to an external climatology (ideally from a data set for a constant climate outside the period of experimentation.) The reason that 3 years should be taken out for the systematic error correction (SEC) is that one can show analytically that the correlation does not change upon

Mdl 4	anomaly	Obs	anomaly	year	SEC	Random Years
25.5	.9	26.8	-.4	1981	-2.62	(1985 1989)
25.9	1.3	28.1	.9	1982	-2.62	(2000 1989)
23.8	-.9	27.1	-.1	1983	-2.46	(1990 1998)
23.5	-1.3	26.7	-.5	1984	-2.44	(1993 1981)
24.1	-.8	26.7	-.5	1985	-2.32	(1992 1995)
26.0	1.4	27.4	.2	1986	-2.56	(1999 1987)
26.6	2.0	28.8	1.6	1987	-2.63	(1996 1989)
23.6	-1.1	25.6	-1.6	1988	-2.50	(1989 1995)
26.2	1.5	26.7	-.5	1989	-2.48	(1983 1992)
25.8	1.1	27.3	.1	1990	-2.54	(1985 2000)
23.5	-1.2	27.9	.7	1991	-2.42	(1990 2001)
24.4	-.3	27.5	.4	1992	-2.49	(1996 2001)
24.4	-.5	27.6	.4	1993	-2.32	(1985 1995)
23.5	-1.3	27.3	.1	1994	-2.38	(1989 1991)
22.9	-1.8	27.0	-.2	1995	-2.48	(1986 1996)
25.6	.9	27.1	-.1	1996	-2.45	(1991 1990)
25.8	1.0	28.9	1.7	1997	-2.36	(1991 1990)
23.4	-1.4	25.9	-1.2	1998	-2.37	(1991 1988)
24.5	-.3	26.3	-.8	1999	-2.42	(2001 1995)
25.0	.2	26.7	-.5	2000	-2.41	(2001 1991)
25.2	.5	27.3	.1	2001	-2.50	(1998 1999)
24.7	.0	27.2	.0	all	-2.45	

Table 1 Shown in column 1 are June temperatures for 1981-2001 (top to bottom) in the Nino34 area as predicted at a lead of 5 months by one of the Demeter models (model#4) which has its initial states in January. The observed SST is shown in column 3. The anomalies in columns 2 and 4 are wrt to the 21 year mean of model and observed data respectively. The bottom line shows 21 year averages. Column 6 shows the systematic error correction (SEC) that would be applied to the year in column 5. Columns 7&8 are two randomly selected years also withheld in calculating the recommended CV3RE SEC.

taking out just 1 year, i.e. CV1 does not do anything. The number of elements withheld being odd (as a convenience), three would thus be the minimum. Typically that would be three successive years as a block, but here we argue that the three removed should be a) the test year, and b) two additional randomly chosen years.

2. An example of systematic error correction

Table 1 provides details of an example. Shown in column 1 are June temperatures for 1981-2001 (top to bottom) in the Nino34 area as predicted at a lead of 5 months by one of the Demeter models (“Model #4”) which has its initial states in January. The observed SST is shown in column 3. The anomalies in columns 2 and 4 are wrt to the 21 year mean of model and observed data respectively. The bottom line shows 21 year averages. Column 6 shows SEC that would be applied to the year in column 5. Columns 7&8 are two randomly selected years also withheld in calculating the recommended SEC.

Clearly model 4 needs systematic error correction badly, since it is about 2.5°C too cold. This is a large error in the mean given that anomalies are rarely larger than 1.5. However, it would be wrong to assume that we know $SEC = -2.45$ with such certainty so as to apply it to all cases in the sample of 21 - this would be the full sample dependent data approach. If one withholds each year in turn (in the hopes of creating an independent year), plus two more years chosen at random, and calculates the difference in the mean of forecast and observation over 18 cases, one finds SEC to vary somewhat but not greatly, from -2.32 to -2.63 to be specific. Fortunately, the forecast still improves greatly as a result of applying a variable SEC, but not as much as, seemingly, when applying a constant $SEC = -2.45$. It is more correct to say that the dependent data case ($N=21$) over-estimates skill, and we have a professional duty to calculate an estimate that will hold up in true real-time. As shown in Fig.1 the skill, as measured by correlation, is around three points lower than in the dependent data result for each of the 9 models on the left in Fig.1 considered by Pena and Van den Dool (2008).

That the year for which forecast accuracy is tested should not be included in the SEC determination is easily seen in the extreme for $N=1$ – that would make the forecast perfect in a misguided way. But even for $N=21$ the test case has a noticeable impact, because of “compensation” effects that are known to affect CV. For instance, in 1987, see Table 1, the forecast and observation are ‘only’ -2.2°C apart and including this case keeps the SEC at -2.45, whereas excluding it makes it -2.63. The opposite happens in 1985 and 1993, two years that feature forecast errors larger than average. Using three elements dilutes the compensation effect. In section 3 we will see a more complicated compensation effect.

In the next section we argue again that three should be taken out, but for a very different reason.

3. Degeneracy in regression

In earlier work we found highly negative correlation in CV applied to forecasts based on regression schemes, where a zero correlation would have been more reasonable. This feature was ultimately explained in Barnston and van den Dool (1993). Fig. 2, reproduced from that paper shows a synthetic data case. We

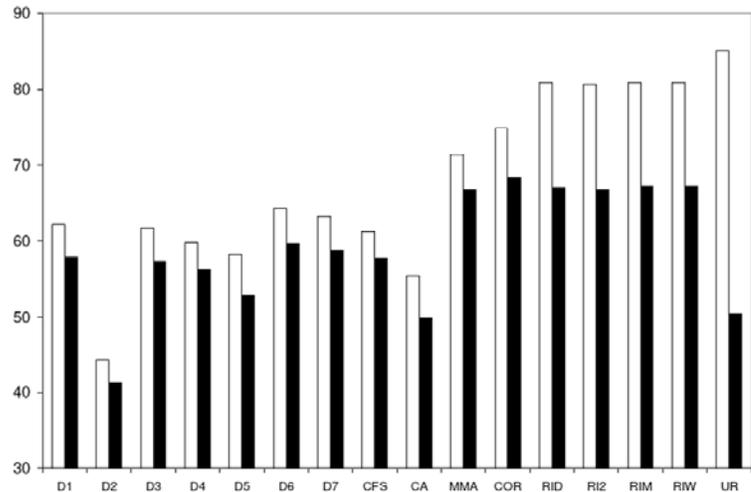


Fig. 1 Anomaly pattern correlation of systematic error corrected monthly SST over the tropical Pacific domain, averaged for all leads and initial months based on the 21 yr of data in the hindcasts (empty bars) and after 3 yr random cross validation (dark bars). The consolidation is done gridpoint-wise, which can be improved upon by increasing effective sample size. On the left the seven Demeter model, the CFS and CA. On the right seven entries for various MME approaches. See Pena and van den Dool (2008).

generated pairs of correlated forecast, observation data, varying the correlation along the x-axis from 0 to 1. We then did a CV-1 approach to calculate the correlation from limited data (32 pairs). When the correlation is large CV-1 functions OK and only shows some normal ‘shrinkage’. But when the intended correlation is small, between 0 and 0.2 in this case, the outcome of CV-1 is a disaster. One can get a perfect -1 correlation. This happens because of compensation effects at the covariance level (in section 2 we had compensation in the mean). Suppose we have zero correlation on the full sample between forecast and observation, and thus also zero covariance. When we leave out 1 pair, which happens to co-vary by chance positively, the remaining N-1 pairs have, by necessity, a negative covariance in the mean. Thus a regression forecast based on the N-1 will be opposite to what is observed in the one case left out, thus leading to high negative correlation.

This can happen in real life regression forecasts. For instance Nino34 correlates with seasonal temperature over the US, but with opposite sign in the NW and the SE US. Along the broad band of zero and small correlation, presumably the nodal line of a teleconnection pattern, the CV-1 score of a regression forecast is highly negative. Here we get punished for our good intentions. The solution, aside from waiting forever for more years, is to take out more than 1. For instance when taking out the test year as well as two more years, the compensation effect is obviously diluted. Choosing two more years at random (as opposed to a block of three, with the test year in the middle) is better because the serial correlation (caused by climate change among other things) violates the assumption of independence.

This above discussion applies to the multi-model ensemble approach because the MME is a linear combination of several forecasts, with weights derived from a limited data set as per regression. We should apply CV3RE, and we can fold the CV for SEC into the CV required for the weights (the regression aspect) into one single procedure. The seven entries on the right hand of Fig.1 are MME by different schemes subjected to CV3RE. The various ridge regression approaches fare much better under CV than an unconstrained regression (UR).

4. Conclusion

We recommend as cross-validation procedure something called CV3RE, where CV is cross-validation, 3 means three years left out, R refers to the random choice of two of the three years left out, and E refers to an external climatology (ideally from a data set for a constant climate outside the period of experimentation.). We have not laid out the case for the external climatology in this short write-up, but this aspect also helps stabilize the answers one gets. While we believe CV3RE is appropriate for the multi-model ensemble it may also be a good strategy in many other situations. However, each problem requires some deliberations of its own, and a general theory/algorithm for CV appears elusive (to me).

References

- Barnston, A.G., and H.M. van den Dool, 1993: A Degeneracy in Cross-Validated Skill in Regression-based Forecasts. *J. Climate*, **6**, 963–977.
- Peña, M., and H. van den Dool, 2008: Consolidation of Multimodel Forecasts by Ridge Regression: Application to Pacific Sea Surface Temperature. *J. Climate*, **21**, 6521–6538.

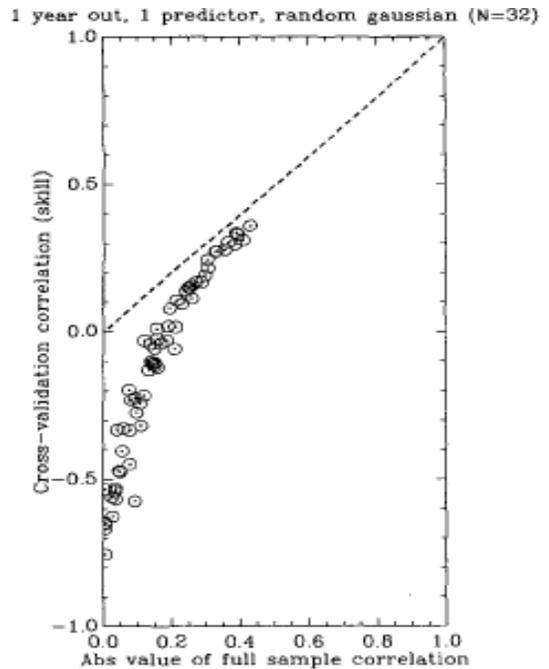


Fig. 2 The correlation (Y axis) calculated as per CV-1 from synthetic data generated by computer with a known correlation (X – axis). For instance when generating paired data with 0.3 correlation the CV-1 procedure (applied to 32 pairs) returns an estimate for the correlation around 0.1.