Geophysica (2010), 46(1–2), 47–58

Estimating Optimal Weights for Combinations of Multiple Forecasts

Jussi Leinonen and Kimmo K. Kahma

Finnish Meteorological Institute, P.O.Box. 503, FI-00101 Helsinki, Finland

(Received:April 2009; Accepted: March 2010)

Abstract

The combining of multiple models is a technique used in forecasting to improve results over individual models. In this paper, we describe a method to produce an optimal composite forecast by decorrelating the original source data, similar to Principal Component Analysis, and combining the decorrelated components with observed data in an optimal manner. As a test of the performance of the method, we used water level forecasts of varying quality from the Finnish Baltic Sea coast, together with tidegauge measurements. The tests show a 5-25% improvement of the forecast error over the best original forecast. The method is robust and suitable for operational use to produce forecasts without human intervention.

Key words: decorrelation, multi-model forecast, learning forecast, principal component analysis

1. Introduction

The mutually cancelling effect of various independent error sources on different forecasts of the same phenomenon gives rise to the ensemble property: the results from different models describing the same phenomenon, affected to various degrees by unrelated error sources, become distributed around the true value of the forecast variable. This variable could be almost any quantitative and measurable value.

Assuming that there are a large number of different, independent factors contributing to the error in each forecast, the central limit theorem would suggest that the various forecasts should be normally distributed around their mean value, thus the sample mean being the best estimate of the true value. This is the basis of multi-model forecasting, on which there exists plenty of previous research (see e.g. *Wandishin et al*., 2001; *Hagedorn et al*., 2005).

In practice, the assumptions of normal distribution and sample mean estimates are dubious at best. Relying on the simple mean may or may not improve the result over the individual forecasts when they are of different quality and the error sources are related to some degree, as indeed usually is the case. This notion suggests that a method which evaluates the quality of the various forecasts and takes the quality into account when building a composite forecast would be a useful tool in improving the quality of forecasts, as noted by *Krishnamurti et al*. (2000).

In this paper, we describe an optimal averaging method for forecasts, making use of data whitening to attempt to extract hidden variables from the forecasts, then using the obtained variables to build a composite forecast. For the sake of clarity, we assume that the forecast and measured variables will be in the form of time series, which is usually the case, but the method can be easily used for prediction over spatial (or any other) variables, as well as interpolation of missing data.

2. *Motivation*

For m different models describing the variable of interest, and n samples from each, let $\mathbf{X}$ be a $m \times n$ matrix containing a forecast time series on every row, each scaled to zero mean and unit variance and linearly independent of each other. Additionally, let $\mathbf{o}$ be a vector of length n, also scaled to zero mean and unit variance, containing the observed variable corresponding to the rows of $\mathbf{X}$.

In an attempt to extract information found in independent "hidden" variables that affect the errors in various forecasts to different degrees, we perform a transformation $\mathbf{S} = \mathbf{TX}$ such that, using the common $\mathbf{R}^n$ inner product notation (for column vectors)

$$< \mathbf{x}, \mathbf{y} >= \mathbf{x}^T \mathbf{y} = \sum_k x_k y_k \tag{1}$$

it holds (within numerical precision) that

$$\frac{1}{n} < \mathbf{S}_i, \mathbf{S}_j >= \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \Rightarrow \frac{1}{n}\mathbf{S}\mathbf{S}^T = \mathbf{I}. \tag{2}$$

Such a transformation is achieved with

$$\mathbf{T} = \mathbf{C}^{-1/2} \tag{3}$$

$$\mathbf{C} = n^{-1}\mathbf{X}\mathbf{X}^{\mathbf{T}} \tag{4}$$

because then, as the covariance matrix $\mathbf{C}$ is always symmetric (see below),

$$\frac{1}{n}\mathbf{S}\mathbf{S}^T = \frac{1}{n}\mathbf{T}\mathbf{X}\mathbf{X}^T\mathbf{T}^T = \mathbf{C}^{-1/2}\mathbf{C}\mathbf{C}^{-1/2} = \mathbf{I}. \tag{5}$$

$\mathbf{C}^{-1/2}$ can be computed from the eigendecomposition

$$\mathbf{C} = \mathbf{V}\Lambda\mathbf{V}^{-1} \tag{6}$$

as

$$\mathbf{C}^{-1/2} = \mathbf{V}\Lambda^{-1/2}\mathbf{V}^{-1} \tag{7}$$

which is simple to compute as Λ is diagonal.

The covariance matrix **C** is symmetric as

$$\mathbf{C}^T = (\mathbf{X}\mathbf{X}^T)^T = \mathbf{X}\mathbf{X}^T = \mathbf{C} \tag{8}$$

and positive semi-denite, since by the denition of inner product, for any **y**

$$\mathbf{y}^T\mathbf{C}\mathbf{y} = \mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{y})^T(\mathbf{X}^T\mathbf{y}) =< \mathbf{X}^T\mathbf{y}, \mathbf{X}^T\mathbf{y} >\geq 0. \tag{9}$$

Since we demanded rows of **X** to be linearly independent, the covariance matrix **C** is nonsingular and thus positive-denite. Therefore its eigenvalues are positive (*Pearson*, 1974) and $\mathbf{C}^{-1/2}$ is guaranteed to exist and be real.

The above procedure is well known and often referred to as "whitening" the data. It is essentially the Principal Component Analysis (*Haykin*, 1999) without dimensionality reduction. A dimensionality reduction step could be used for the method described in this article, and has the known benefit of reducing noise and preventing overlearning, but it is not generally necessary as the weight determination method already reduces the influence of less significant components, and dimension reduction would introduce new free parameters. The number of dimensions in the input data is also often small, further reducing the need to use dimension reduction.

3. Derivation

In order to obtain a composite forecast, we need to estimate an optimal prediction using the uncorrelated components. Our goal is to determine a weight vector **w** such that the linear combination of the components, $\mathbf{w}^T\mathbf{S}$, optimally predicts the observed values **o**. As a measure of optimality, we use the square of the correlation $c(\mathbf{w})^2 = \mathrm{Corr}(\mathbf{w}^T\mathbf{S}, \mathbf{o})^2$.

Since mean values of the rows of X are zero, so is that of $\mathbf{w}^T\mathbf{S}$, as regardless of **w**, as

$$\mathrm{E}(\mathbf{w}^T\mathbf{S}) = \mathrm{E}(\mathbf{w}^T\mathbf{T}\mathbf{X}) = \frac{1}{n}\sum_i\sum_j w_i T_{ij}\sum_k X_{jk} = 0. \tag{10}$$

The variance is

$$\mathrm{Var}(\mathbf{w}^T\mathbf{S}) = \frac{1}{n} < (\mathbf{w}^T\mathbf{S})^T, (\mathbf{w}^T\mathbf{S})^T >= \frac{1}{n}\mathbf{w}^T\mathbf{S}(\mathbf{w}^T\mathbf{S})^T$$

$$= \frac{1}{n}\mathbf{w}^T\mathbf{S}\mathbf{S}^T\mathbf{w} = \mathbf{w}^T\mathbf{w} =< \mathbf{w}, \mathbf{w} > \tag{11}$$

and thus we can define $\mathbf{w}^T\mathbf{S}$ to have unit variance by requiring that

$$< \mathbf{w}, \mathbf{w} >= 1. \tag{12}$$

The zero mean and unit variance guarantee that the correlation is equal to the covariance $\mathrm{Cov}(\mathbf{w}^T\mathbf{S}, \mathbf{o})$. Then we have

$$\begin{aligned} c(\mathbf{w}) = \mathrm{Corr}(\mathbf{w}^T\mathbf{S}, \mathbf{o}) &= \frac{1}{n} < (\mathbf{w}^T\mathbf{S})^T, \mathbf{o}^T > \\ &= \frac{1}{n}\mathbf{w}^T\mathbf{S}\mathbf{o}^T \\ &= \frac{1}{n} < \mathbf{w}, \mathbf{S}\mathbf{o}^T > . \end{aligned} \tag{13}$$

We can now show that the optimal choice is

$$\arg\max c(\mathbf{w}) = \arg\max c(\mathbf{w})^2 = \frac{\mathbf{S}\mathbf{o}^T}{\sqrt{< \mathbf{S}\mathbf{o}^T, \mathbf{S}\mathbf{o}^T >}} = \mathbf{S}'\mathbf{o}^T. \tag{14}$$

Firstly, it satisfies the constraint $< \mathbf{w}, \mathbf{w} >= 1$ as

$$\begin{aligned} < \mathbf{S}'\mathbf{o}^T, \mathbf{S}'\mathbf{o}^T > &= \frac{< \mathbf{S}\mathbf{o}^T, \mathbf{S}\mathbf{o}^T >}{< \mathbf{S}\mathbf{o}^T, \mathbf{S}\mathbf{o}^T >} \\ &= 1. \end{aligned} \tag{15}$$

Secondly, it can be shown that it is indeed the optimal choice: let $\tilde{\mathbf{w}}$ be any vector of size *m* such that

$$< \tilde{\mathbf{w}}, \tilde{\mathbf{w}} > = 1 \tag{16}$$

$$\tilde{\mathbf{w}} \neq \mathbf{S}'\mathbf{o}^T. \tag{17}$$

Then we have from the Cauchy-Schwarz inequality that

$$\begin{aligned} < \tilde{\mathbf{w}}, \mathbf{S}'\mathbf{o}^T >^2 &\leq < \tilde{\mathbf{w}}, \tilde{\mathbf{w}} >< \mathbf{S}'\mathbf{o}^T, \mathbf{S}'\mathbf{o}^T > \\ &= 1 \\ &= < \mathbf{S}'\mathbf{o}^T, \mathbf{S}'\mathbf{o}^T >^2 \end{aligned} \tag{18}$$

$$= < \mathbf{w}, \mathbf{S}'\mathbf{o}^T >^2 . \tag{19}$$

Thus the optimal weight coefficients w_i are the correlations of the corresponding component $\mathbf{S}_i$ and the observation $\mathbf{o}$. With these coefficient known, the optimal forecast estimate is then obtained by re-scaling the estimate $\mathbf{w}^T\mathbf{S}$ to the appropriate mean and variance (or standard deviation). If observed data are available for the entire time range of the estimate, we can simply use the observed mean and variance. In the next section,

we discuss the more realistic case, in which we attempt to estimate the optimal forecast for a period where the measurement is unavailable.

One should note that the approach described above very closely resembles the estimation of a function in an incomplete orthogonal basis.

4. Learning prediction

To be useful as a forecast tool, the method must be able to be used when no measured time series is available. To achieve this, we determine the optimal coefficients for a period where the true value is known, and then apply those coefficients for periods without a measurement. For purposes of evaluating the learning ability, we divided our dataset in two parts, called (per the convention in learning systems) the teaching set and the test set. In the teaching set, we use both the forecasts and the observation to "teach" the proper coefficients and transformation matrices to the system; in the test set, the observation is only used to evaluate the performance and learning ability of the method.

Formally, we proceed as follows:

1. Determine the whitening transformation and optimal coefficients, as described in the previous sections, for the teaching dataset $\mathbf{X}_1$, to obtain the transformation matrix $\mathbf{C}_1^{-1/2}$ and the weight coefficients $\mathbf{w}_1$.
2. Apply the transformation matrix to the test dataset $\mathbf{X}_2$ to obtain

$$\mathbf{S}_{12} = \mathbf{C}_1^{-1/2}\mathbf{X}_2. \tag{20}$$

3. Estimate the observed value for the test set as

$$\mathbf{e}_{12} = \mathbf{w}_1^T\mathbf{S}_{12}. \tag{21}$$

As $\mathbf{e}_{12}$ is still scaled to non-physical mean and variance, we need to rescale it to obtain the values on a proper scale. For both the mean and the standard deviation, we have used a weighted average of the corresponding values for the different forecasts, using as the weights the correlation of the forecast with the measurement in the teaching set $\mathbf{X_1o}$, normalised such that the sum of the weights is 1.

5. Experiments

We have used the learning forecast method on water level forecasts for the coast of Finland. Three Baltic Sea water level models were used, originating from different sources: the Swedish Meteorological and Hydrographical Institute (SMHI) model, the German Federal Maritime and Hydrographic Agency (Bundesamt für Seeschifffahrt und Hydrographie, BSH) model and the Wetehinen model being developed at the Finnish Meteorological Institute. The SMHI model (*Funkquist*, 2001) and the BSH model are both 3D hydrodynamic models based on the same code from BSH (*Dick*, 2001). While the two models have developed further, and the codes are no longer exactly identical,

the main differences are in the operational implementations. The Wetehinen model, on the other hand, is a less sophisticated 2D model and has a much higher forecast error than the other two.

The data from each forecast was level-corrected as a pre-processing step to compensate for level drifting and different reference water levels. This was done by adjusting the data at each sample by the difference of the mean values of the forecast and observed values for the previous week. This aims to make the weekly averages equal for forecasts and measurements without evaluating future data (which would not be available for an operational forecast).

Our dataset consisted of measured and forecast water level values for the coast of Finland from June to December, 2007. The number of data points in the set was $n = 1810$. In operational use, the method would likely be taught with data from the previous few months of forecasts and measurements. For testing purposes, the dataset was divided into two halves of equal size (the earlier half called A and the later one called B). We alternated them as the teaching and test sets and evaluated the results. As a measure of forecast error, we used the root-mean-square error (RMSE)

$$e_{\mathrm{RMS}} = \sqrt{\frac{1}{n}\sum_{k}^{n}(x_{\mathrm{fc},k} - x_{\mathrm{obs},k})^2} \tag{22}$$

where x_{fc} and x_{obs} are the forecast and observed data, correspondingly. We used the RMSE here because it provides simple, physically meaningful values, however, one should note that since our method is designed using the correlation as the measure of forecast error, the final estimate may not always be exactly optimal in the RMSE sense (although a correlation of 1 does imply an RMSE of 0).

In Table 1, summarised in Figure 1, we present the errors from different combinations of teaching and testing sets at different sites. For comparison, we also show the errors of the original forecasts. We see that for both selections of the test set, the combined forecast always improves the quality of the forecast over the best individual source, despite there often being considerable differences in the errors of the originals. Thus we see that useful information that improves the combined forecast can be extracted even from the poorer quality forecasts. Figure 2 is an example of the behaviour of the combined estimate.

We also compare the performance of the method to an intuitive alternative, a weighted average that uses as the weights the normalised inverse values of the RMSE of the corresponding forecasts. This approach also tends to outperform the best individual forecast. Our method produces significantly better results than the weighted average in two cases (Kemi and Kaskinen), and does about equally well in the two others.

Table 1. The RMS errors (cm) for various forecast types and locations on the coast of Finland.

Forecast	Kemi	Kaskinen	Föglö	Helsinki
A teaching and testing	4.10	2.61	2.18	3.10
B teaching and testing	5.15	3.33	2.28	3.57
B teaching, A testing	4.17	2.64	2.37	3.24
A teaching, B testing	5.24	3.33	2.42	3.66
A+B teaching and testing	4.70	3.02	2.30	3.37
Weighted average (A+B)	4.79	3.14	2.29	3.38
SMHI (A)	4.51	2.85	2.55	4.14
SMHI (B)	5.86	3.64	3.04	4.87
SMHI (A+B)	5.23	3.27	2.80	4.52
BSH (A)	5.87	3.47	2.77	4.25
BSH (B)	8.42	4.57	3.45	4.87
BSH (A+B)	7.26	4.06	3.13	4.57
Wetehinen (A)	7.95	5.46	3.82	6.44
Wetehinen (B)	11.63	6.15	2.98	5.59
Wetehinen (A+B)	9.96	5.81	3.42	6.03

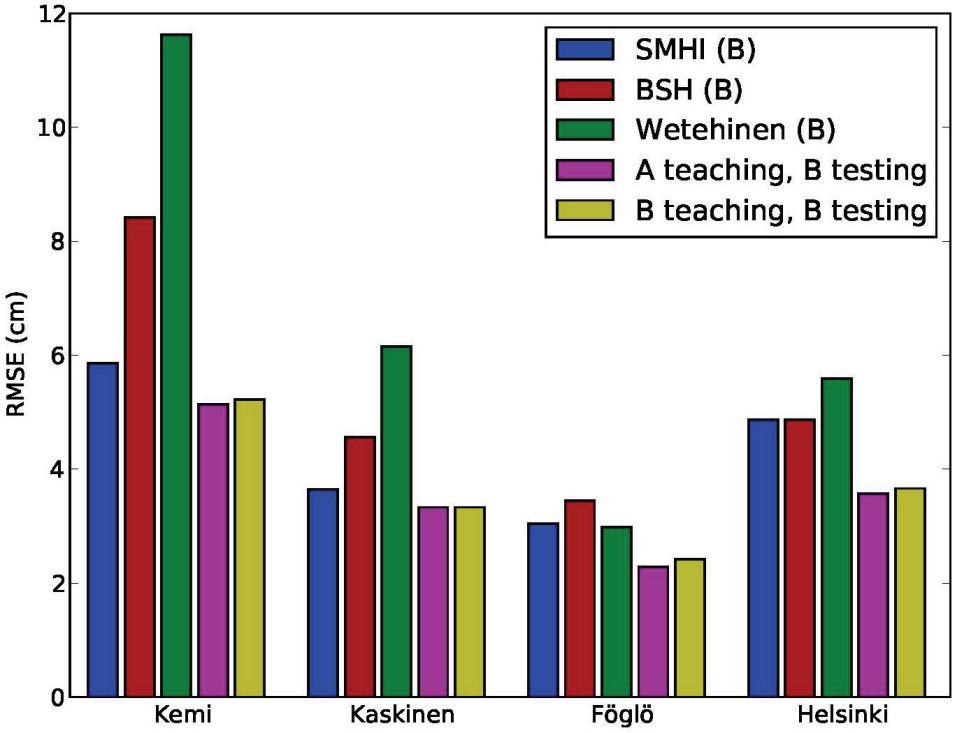

Fig. 1. Summary of the results for the RMS error of various methods for the B set. Bars are in the same order as in the plot legend.

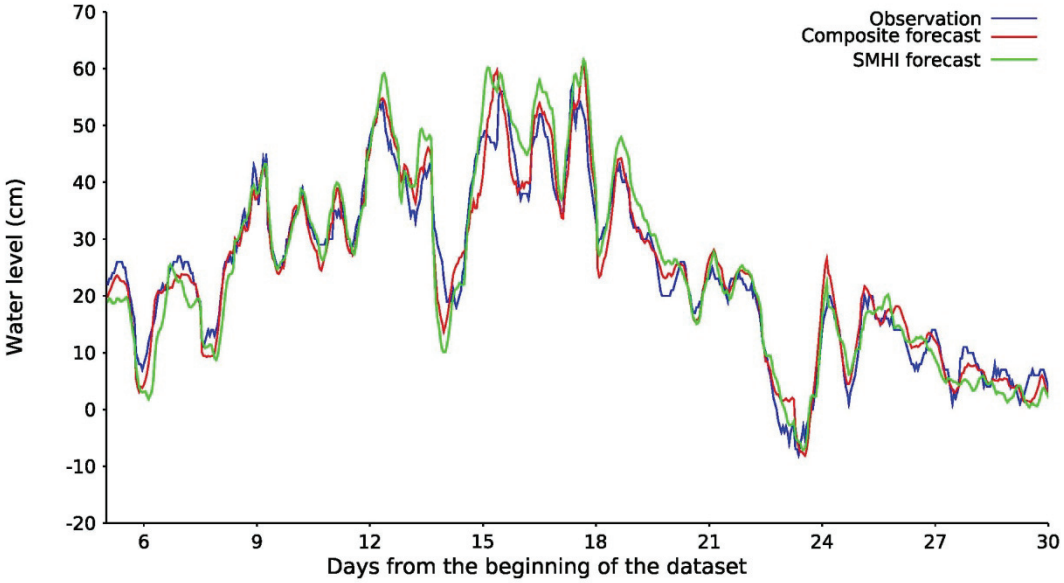

Fig. 2. The composite forecast and the best original (SMHI) forecast compared to the measured water level at the Helsinki site.

Tables 2 and 3 show the same analysis as Table 1, but with the RMSE evaluated only for the 10 % of highest and lowest water levels, respectively. For the extremes, the result is still usually (though not always) improved. Since we are sampling the error estimate from a limited subset of the data, there is no theoretical guarantee of optimality for the teaching set, as can be seen from the case of high water in Kaskinen, where using the A set for teaching and the B set for testing gives a better result than using the B set for both teaching and testing. An alternate solution would have been to use only the extreme data for teaching, but experiments with this indicated that it leads to poor results when different teaching and test sets are used. This is probably due to the small size of the teaching set (about 220 data points) when only maxima and minima are used, though it may also indicate that using only the extremes is not sufficient to characterize the behaviour of the various sources. It is notable and relevant to applications that the advantage of our method over the weighted average is much more pronounced at the extrema than it is in the full dataset.

The measured water levels and the corresponding forecasts at the highest and lowest levels (measured from peak maxima and minima, respectively) at each site are found in Table 4. The results are somewhat mixed, and it can be seen that when the error of the source forecasts is large, so is that of the composite. This is not surprising: if the sources do not contain the information from which to predict the correct water level, there is little room for computational improvement. Another source that can contribute to errors near peak levels is the difference in the peak timings for various models; the composite method does not attempt to correct for such timing errors. Figure 3 shows that for the typical extremal cases these errors are usually not very large compared to the width of the peaks, although the shapes of the peaks differ.

Table 2. The RMS errors (cm) for various forecast types and locations on the coast of Finland for the 10 % of highest water levels.

Forecast	Kemi	Kaskinen	Föglö	Helsinki
A teaching and testing	7.03	3.59	2.34	4.22
B teaching and testing	8.29	3.67	2.29	4.74
B teaching, A testing	7.09	3.74	2.66	4.57
A teaching, B testing	8.49	3.51	2.48	4.93
A+B teaching and testing	7.61	3.88	2.47	4.78
Weighted average (A+B)	7.81	3.98	2.46	4.86
SMHI (A)	7.81	3.70	2.61	4.67
SMHI (B)	9.88	4.85	4.38	6.87
BSH (A)	8.00	4.95	2.97	6.40
BSH (B)	14.53	5.19	3.79	6.64
Wetehinen (A)	12.41	6.27	4.14	8.95
Wetehinen (B)	16.97	5.06	2.85	6.45

Table 3. The RMS errors (cm) for various forecast types and locations on the coast of Finland for the 10 % of lowest water levels.

Forecast	Kemi	Kaskinen	Föglö	Helsinki
A teaching and testing	4.55	2.37	2.36	2.71
B teaching and testing	6.05	3.23	2.22	3.17
B teaching, A testing	5.04	2.67	3.13	2.79
A teaching, B testing	6.13	3.27	2.09	3.42
A+B teaching and testing	5.28	2.49	2.28	3.10
Weighted average (A+B)	5.43	2.82	2.53	3.38
SMHI (A)	5.04	2.62	3.22	3.35
SMHI (B)	5.79	3.56	2.59	4.43
BSH (A)	6.58	3.09	2.99	3.97
BSH (B)	11.67	4.90	2.64	4.22
Wetehinen (A)	6.66	5.19	4.52	6.32
Wetehinen (B)	15.64	6.84	3.36	6.81

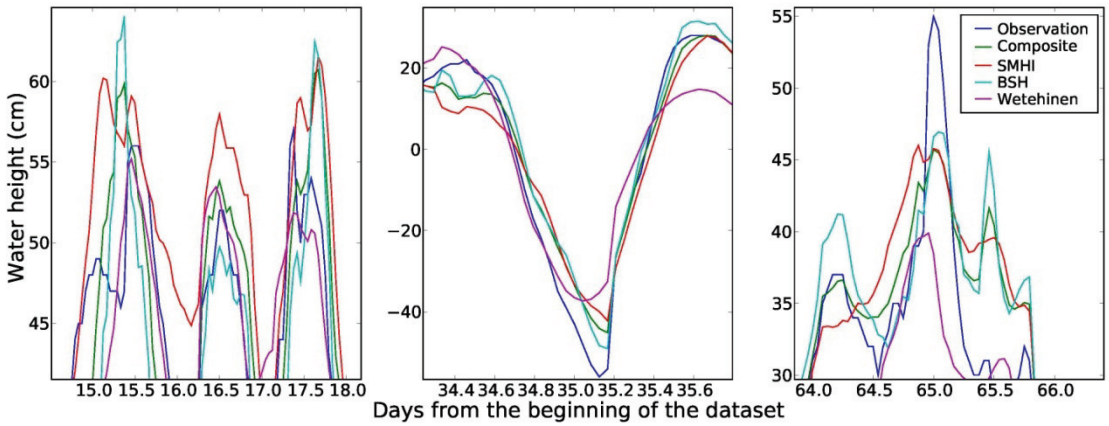

Fig. 3. The behaviour of the different forecasts at extremal values of the water height in Helsinki.

Table 4. The highest and lowest measured water levels (cm) for each site and the corresponding forecasts.

Case	Measurement	Composite	SMHI
Kemi (high)	86	80	82
Kemi (low)	-62	-48	-47
Kaskinen (high)	47	44	46
Kaskinen (low)	-35	-31	-28
Föglö (high)	45	49	52
Föglö (low)	-13	-14	-17
Helsinki (high)	71	63	66
Helsinki (low)	-56	-45	-42

As the composite forecast is computed as $\mathbf{w}^T\mathbf{S} = \mathbf{w}^T\mathbf{T}\mathbf{X}$, the relative importance of the normalised models can be seen from the vector $\mathbf{w}^T\mathbf{T}$. These are shown for the full dataset at each location in Table 5. Since the result is a weighted average of the normalized forecasts, there tends to be some smoothing of the values. However, negative weights may also be assigned to the forecasts, which shows that the difference of forecasts can also contain relevant information. Interestingly, this happens in the case of Kaskinen, where the composite method also does best compared to the weighted average. We also show the eigenvalues of the correlation matrix in Table 6. It shows that largest eigenvalue is several orders of magnitude larger than the others in all cases, which indicates that the less significant components have importance mainly in fine-tuning the result.

Table 5. The model weights $\mathbf{w}^T\mathbf{T}$ for the full dataset A+B at each location, normalized to a sum of 1.

	Kemi	Kaskinen	Föglö	Helsinki
SMHI	0.68	0.69	0.40	0.36
BSH	0.30	0.32	0.33	0.46
Wetehinen	0.03	-0.01	0.27	0.18

Table 6. The eigenvalue vectors Diag(Λ) of the correlation matrix **C** for each location.

Kemi	Kaskinen	Föglö	Helsinki
$\begin{pmatrix} 13.0 \\ 0.350 \\ 0.170 \end{pmatrix}$	$\begin{pmatrix} 4.66 \\ 0.0936 \\ 0.0460 \end{pmatrix}$	$\begin{pmatrix} 3.82 \\ 0.0608 \\ 0.0329 \end{pmatrix}$	$\begin{pmatrix} 4.76 \\ 0.179 \\ 0.0787 \end{pmatrix}$

6. Conclusions

We have described an optimal method to estimate coefficients to produce composite forecasts from several independent sources. To achieve an optimal forecast, the method combines uncorrelated components of the original forecasts by maximizing the correlation of the result with a previously known measurement. Having been taught in this manner, the predictor can be used to make actual forecasts for the future or other periods where the observed value is not available.

We used the predictor to create forecasts for water level at the coasts of Finland. Three models were used in the test: Two were nearly identical with main differences in the operational implementation. The third one was inferior compared with the two others. One could think that this kind of combination would not improve the accuracy or to be at best as good as the most accurate model. Our results show the opposite. For every test case, the composite forecast provides a considerable (5–25 %) improvement over the best individual forecast. This demonstrates the ability of the method to use information from multiple forecasts of different overall quality.

It should also be noted that the result is achieved with a fairly straightforward, linear algorithm. The advantage of this is its small number of free parameters, which prevents overlearning of the teaching data. This makes the method especially robust in operational use, the only concern being the requirement of linear independence, which usually holds in practice, but should be evaluated before running the analysis. Extending the method to use non-linear prediction such as the multilayer perceptron (*Haykin*, 1999) should be possible, though it may come at the expense of generalization ability.

References

Dick, S., E. Kleine, S.H. Müller-Navarra, H. Klein, H. Komo, 2001. Operational circulation model of BSH (BSHcmod) – model description and validation. *Berichte des BSH* **29**.

Funkquist, L., 2001. Hiromb: An operational eddy-resolving model for the Baltic Sea. *B. Maritime I. Gdansk* **28** (2), 7–16.

Hagedorn, R., F.J. Doblas-Reyes, T. Palmer, 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting – I. basic concept. *Tellus A* **57** (3), 219–233.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation (2nd Edition). Prentice Hall.

Krishnamurti, T.N., C.M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil and S. Surendran, 2000. Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate* **13** (23), 4196–4216.

Pearson, C. E., 1974. Handbook of Applied Mathematics. Von Nostrand Reinhold Company.

Wandishin, M.S., S.L. Mullen, D.J. Stensrud and H.E. Brooks, 2001. Evaluation of a short-range multimodel ensemble system. *Mon. Weather Rev*. **129** (4), 729–747.