Procedia Computer Science

Volume 66, 2015, Pages 220–227

YSC 2015. 4th International Young Scientists Conference on Computational Science

Dynamic Selection of Ensemble Members in Multi-Model Hydrometeorological Ensemble Forecasting

Alexey V. Krikunov and Sergey V. Kovalchuk

ITMO University, Saint-Petersburg, Russia
alexey.v.krikunov@ya.ru kovalchuk@mail.ifmo.ru

Abstract

Multi-model prediction ensembles show significant ability to improve forecasts. Nevertheless, the set of models in an ensemble is not always optimal. This work proposes a procedure that allows to select dynamically ensemble members for each forecast. Proposed procedure was evaluated for the task of the water level forecasting in the Baltic See. The regression-based estimation of ensemble forecasts errors was used to implement the selection procedure. Improvement of the forecast quality in terms of mean forecast RMS error and mean forecast skill score are demonstrated.

Keywords: ensemble, dynamic ensemble selection, classification, multi-Model Ensemble Forecasting

1 Introduction

Multi-model ensemble prediction systems show convincing ability to improve forecasts in different areas of computational science [1]. There are several methods to combine different forecast sources to one ensemble forecast: weighted mean (building on usual using linear-regression) [2], Bayesian models averaging [3] and others.

The task of selection of the ensemble members from the point of view of decreasing computational complexity was discussed by Raftery et al. [3]. Using a large amount of models in the ensemble can be non-optimal in terms of forecast quality due to multicollinearity of forecasting models in the ensemble [4]. Moreover, the optimal set of models in the ensemble may vary in time (in case of continuous forecasting scenarios). This triggers development procedures of selection of ensemble members that will provide selection of an optimal set of models.

In the area of pattern recognition ensemble-based systems used to combine different classifiers and several methods for dynamic ensemble selection were proposed [5, 6]. Base idea of dynamic classifier ensemble selection (or single classifier selection) finds in training set K samples nearest for current observation which we need to classify. Then only those models are selects, which gave right answer for all K nearest samples in training set. All classifiers have same input in classification time, but forecasts can be based on different input data. Another difference is that in most situations we can't say that forecast was right; we only can say that

 Selection and peer-review under responsibility of the Scientific Programme Committee of YSC 2015

doi:10.1016/j.procs.2015.11.026

it has some forecast error, which can be more or less. We can rank models by forecast error on K nearest training set samples, using different training sets for each forecast source, but it remains an open question how much models should be selected. On the other hand we can't be sure that ranking using different training sets is representative. Due to these differences between the ensembles of classifiers and ensembles of forecasts methods for classifier ensembles selection are not suitable for forecasting.

2 Theoretical background

2.1 Ensemble skill

To evaluate the quality of forecasts a skill score is normally used [7]. A skill score (or prediction skill) express the increase of capability of forecast given by model compared to a reference forecast with respect to some prediction score. The reference usually represents an unskilled or low-skilled forecast. Prediction score shows accuracy of prediction made by model: the lower score shows model, the better forecast gives model. In meteorology to represent an unskilled forecast the three standards are commonly used "chance", "persistence", and "climatology" [7]. In our study we use *climatology* that is a forecast of the long term average of the forecasting value. The skill score is:

$$SS = \frac{Score_{forecast} - Score_{reference}}{Score_{perfect} - Score_{reference}} \tag{1}$$

Where $Score_{forecast}$ is a prediction score for the investigated model, $Score_{reference}$ is the reference score and $Score_{perfect}$ is the best possible value for a given score.

There are different ways to calculate score but the forecast error is usually used for deterministic forecasts of continuous values (e.g water level). Forecasts often makes in form of time series. Most commonly used measures of forecast error for time series are root mean squared error (RMSE) and mean average error (MAE). In this study we use the Root Mean Square Error Skill Score (RMSE-SS) skill score based on RMSE values. A main advantage of RMSE is that it gives more weight to larger errors.

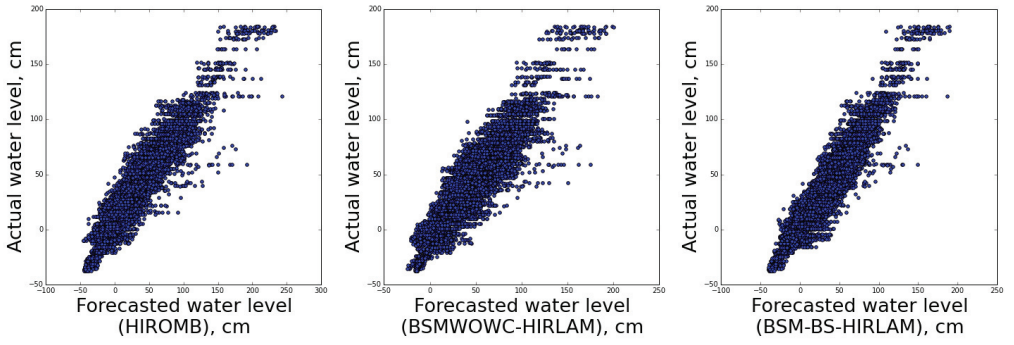
RMSE-SS is defined as:

$$SS = 1 - \frac{RMSE_{forecast}}{RMSE_{reference}} \tag{2}$$

2.2 Ensemble members selections

Multi-model ensembles make forecasts by combining forecasts of different models. Thus the ensemble aims to balance weaknesses of some models with strengths of others. But usually there are uncertainties that cannot be considered by any of forecasting models and as a consequence any of ensembles. The plots in Figure 1 show forecasted and actual values of water levels of Baltic Sea in area of Saint-Petersburg (for details see the Section 4) for three different models. For some models the variance of the forecast error changes depends of the actual value of the level. Heteroscedasticity of the forecasted values may imply that there are dependencies between current water levels and the skill of the forecasting model.

In some cases it is possible to develop a selection procedure that allows us to automatically select the appropriate model or set of models for ensemble building for every state of the modeling system. It can be considered as an evidence of a dependency (often implicit or caused by model limitations) between values of some properties of the modeling system and skill of the



(a) (b) (c)

Figure 1: Forecasted versus actual water level for different models: a)HIROMB, b)BSM-WOWC-HIRLAM and d)BSM-BS-HIRLAM

ensemble or particular model. Construction and calibration of a new ensemble at each forecast time can be a task with a high computational complexity. Instead of that we could take one ensemble from the set of pre-constructed ensembles in accordance with the selected models. If we have N different models we can build 2^N ensembles or 2^{N-1} ensembles if the "ensemble" with only one source – average value (in this case averaged water level – climatology) isnt considered. This set of ensembles allows us to make multi-ensemble forecast similar to multi-model forecasts. Thus N can be quite large; in the study [4] it was shown for an example of 17 models that adding a new model at the ensemble can causes decrease in quality of the ensemble forecast due to multicollinearity of the forecasts data for similar models.

3 Developing a selection procedure

If we cannot choose one model or ensemble of models that will give best results in every situation we can try to choose an ensemble every time we need a forecast. The selection procedure can be based on analyzation of the state of the modeling system and comparing it with ensemble properties.

3.1 Ensemble error prediction

Skill score is inversely proportional to the RMSE of the ensemble forecast. The idea underlying our method is that the ensemble which will show lowest RMSE should be selected to make a forecast in the current situation. As described above, in case of heteroscedaticy there is possible correlation between value of prediction (and real value) and variance of forecast error. Variance in its turn shows how much forecast error can give an ensemble. Correlation between prediction and variance allows estimation of possible forecast error, if the form of dependency between these two values can be found. We build a simple linear model that describes the dependency between the current water level and the ensemble forecast error:

$$E = \alpha + \beta_0 X_0 + \ldots + \beta_n X_n + \epsilon \tag{3}$$

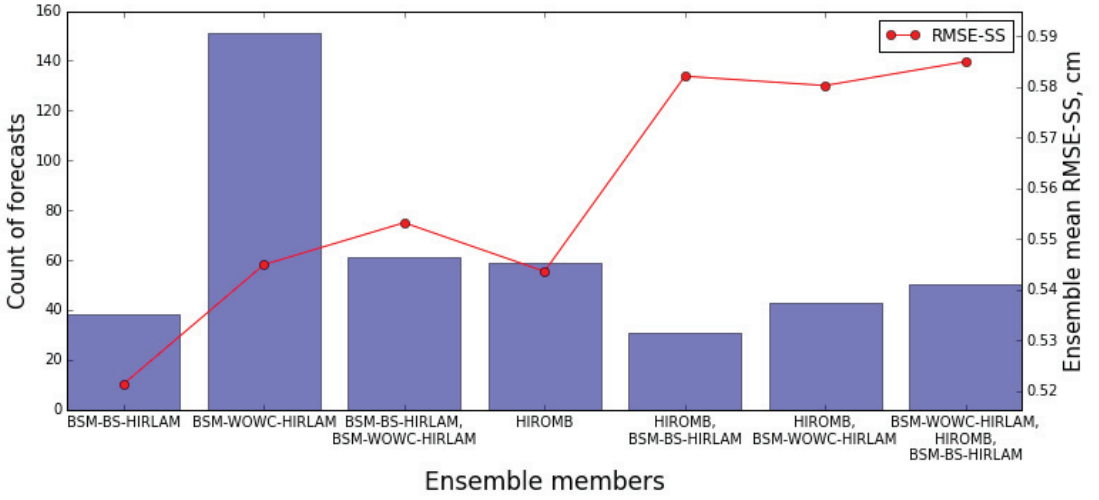


Figure 2: Counts of forecasts when ensemble shows lowest RMS forecast error and respective mean skill score for each ensemble. Ensemble composed of "BSM-WOWC-HIRLAM" model only in most cases has best skill from the set of ensembles, but mean skill score of the full ensemble ("BSM-WOWC-HIRLAM, HIROMB, BSM-BS-HIRLAM") still the best.

Where α and β are regression coefficients, ϵ is error term, and X_n is the level of water n time steps before the forecast time. These linear model parameters are selected for each ensemble using Ordinary Least Squares (OLS). The data that used to build OLS called is training set. Before making a forecast we predict possible forecast error value for each ensemble using these models, and then select the ensemble with smallest predicted error. This selected ensemble is called the *predicted best ensemble*, or *selected ensemble*. The predicted ensemble forecast will be used as result output. Ensemble, which actual forecast RMSE is the least called *best ensemble*.

3.2 Impact of wrong selection

In contrast with common classification tasks in most applications such as diagnosis of diseases or pattern recognition, a wrong selection of the best ensemble has less negative impact on the forecast task, because even a wrong prediction of the best ensemble can a give better or same forecast skill than permanently using of one good calibrated ensemble. E.g. we have one good calibrated ensemble A, and two week ensembles B and C. It is possible that $RMSE_A > RMSE_B > RMSE_C$. Next suppose that our model predicts that ensemble B will show lowest $RMSE$. We select forecast B, that actually is not the best, but still better than forecast of ensemble A which forecast will be used in traditional way multi-model forecasts. That means that we get significant fault only in situations when selected ensemble gives actual forecast error greater than actual forecast error of the *"best-in-average"* ensemble.

4 Experimental study

The study was performed using measurements and models forecasts of the water level in the Baltic Sea. In our study we constructed 8 multi-model ensembles of various combinations of

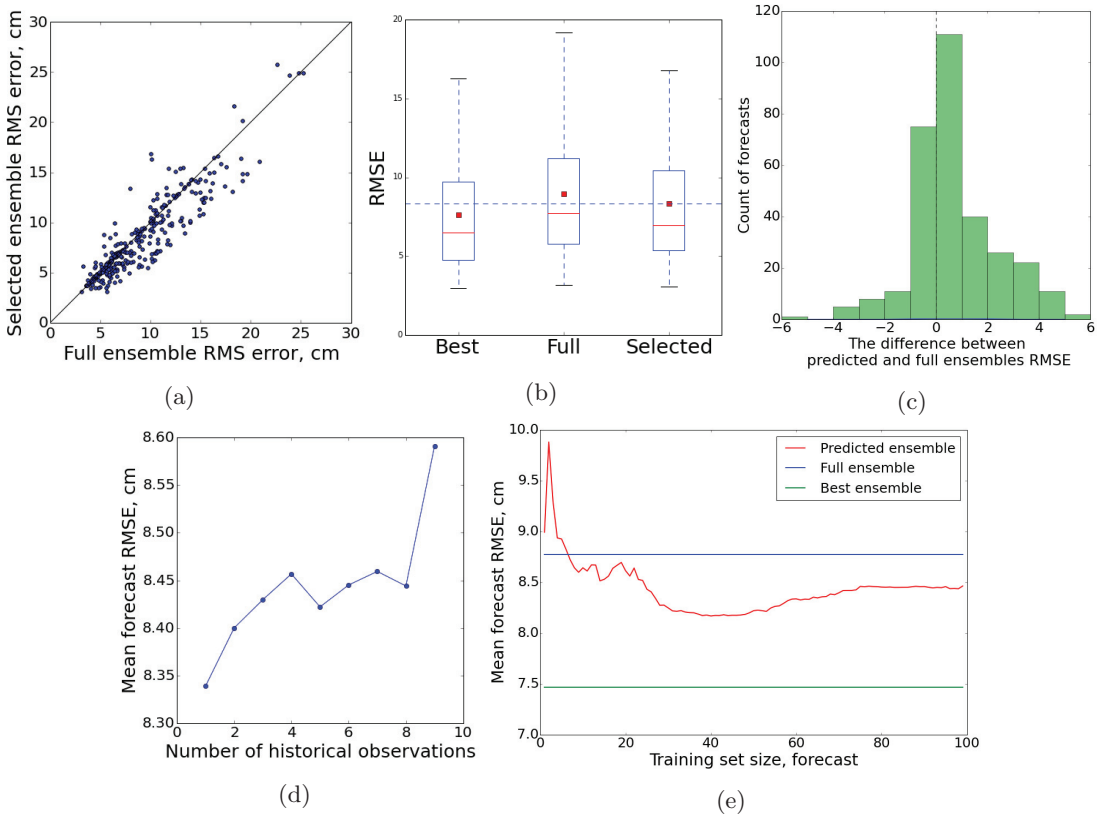


Figure 3: Experimental study on water level ensemble forecasting: 3a) RMSE of predicted to full ensemble; 3b) RMSE distribution for full, best possible and selected ensembles; 3c) Distribution of full and selected ensemble RMSE difference; 3d) Mean forecast RMSE versus number of historical observations; 3e) Mean forecast RMSE versus selection procedure training set size;

the forecast sources and calibrated them using OLS-regression.

As sources results of two runs of the software packages BSM-2010 (Baltic Sea Model) with an external meteorological forecasts source HIRLAM were used. Runs were made with two different execution sets of parameters: with usage of additional data from the spectral wave model SWAN (BSM-WOWC-HIRLAM) and without using SWAN (BSM-BS-HIRLAM). Another source is a water level forecast from external data source - HIgh Resolution Oceanographic model of the Baltic Sea (HIROMB). Each ensemble produces a 60-hours forecast with a time step of 1 hour. Forecasts were made every 6 hours, 434 forecasts in total (covering period from 01.08.2011 to 17.11.2011).

The *full ensemble* (ensemble combined from all three models) shows the lowest mean RMSE on validation data. But even on this small multi-ensemble set full-ensemble do not show best skill for every forecast, still showing best mean skill score (Figure 2).

Our current study was performed on historical data of water level. One part of the historical data was used for ensembles calibration and other part for validation of selection procedure.

Result of evaluation shows that in most cases (67% of all forecasts) the selected ensemble

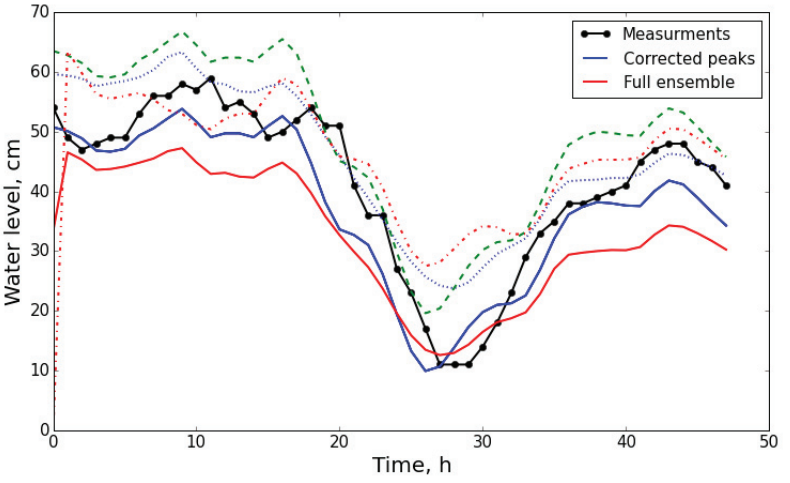


Figure 4: An example of the forecast where selected ensemble was more skillful than full ensemble.

gives less or equal forecast RMSE than the full ensemble and as a result gives lower mean RMSE (Figure 3).

Model described by Equation (3) shows dependency of error from values of water levels for n time steps before forecast. In our case we use only one value of the current water level on the moment of forecast. Using more than one value decreases model accuracy and increases forecast RMSE (Figure 3d).

At the moment of forecast we train our model of forecast error using results of ensembles forecasting in past. Figure 3e shows that only several nearest forecasts in past should be used for model training. Using more data increases the forecast error. Further results were obtained using training set size of 45 forecasts back from the current forecast.

An example of forecast made by models alone, using multi-model ensemble and using multi-ensemble approach is shown in Figure 4. One can see that in this case forecast of the selected ensemble was more skillful than full-ensemble forecast.

5 Discussion

Classification using SVM. The task of selection the best or right ensemble from a set of ensembles according to some values resembles a classification task. We can divide possible states of the modeling system into classes and associate each class with one of the ensembles that shows best forecasting skill for the system in this state. At the moment of forecast we should identify the state of the modeling system and the select appropriate ensemble.

Classification is a common task in machine learning, and the Support Vector Machine (SVM) one of the widely used approaches. SVM is a classification algorithm originally developed by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963 and later extended by number of researcher [8, 9]. This algorithm is based on the idea of construction a dividing hyperplane that has a maximal margin between vectors of two classes. SVM has strong theoretical foundations and empirical successes in wide area of science and engineering.

In our study procedure based on classification using SVM not demonstrated significant

results as selection procedure in contrast with approach based on ensemble error prediction. In the case of SVM classification we got a 2% RMSE increasing in comparison with the results of the full ensemble. It can mean that using of common machine learning algorithm without taking into account specificity of the subject area cannot achieve significant results and that special methods for multi-ensemble forecasting should be developed.

Inaccurate RMSE values. If we make a forecast at a time shorter than the length of the forecasting period, we can only calculate a "partial" forecast RMSE. While this "partial" RMSE was calculated only for the first part of the forecast (that in fact usually have lower deviation from actual value than the "tail" of the forecast) it can be more biased than "full"-RMSE. Error prediction that will be based on the regression results on this data can give "optimistic" result in some cases. It is important that this "biased"-RMSEs cannot be regarded as outliers. Based on this we can assume that using of a method deferent from simple linear regression to error model calibration may give better results. This method should takes into account that different samples in training data may have different degree of trust.

Different metrics of the forecast error. In our current study the RMS error was used as a metrics for scoring. Other popular metrics to measure the distance between time series are mean average error (MAE) and dynamic time wrapping (DTW). Often it is required to involve more complicated statistical [10] or entropy [11] metrics. Still the right selection of metrics requires involvement of domain knowledge.

Features for selection procedure. Only one value level of the water at the moment of forecast was used as a input value for the selection procedure. Next important issue is a search for other properties of the forecasts or ensembles that can be used in selection procedure. And both domain specific knowledge (e.g. knowledge about cyclones and flood connected-knowledge for weather forecasting) and domain-independent knowledge (e.g. spread of the ensemble [12]) should be considered.

Disasters prediction tasks. The ensemble prediction systems are used for disasters prediction tasks such as ensemble flood forecasting [13]. Since error in disaster prediction has higher cost than error simple weather prediction risks of the usage of the proposed method need to be estimated.

6 Conclusion and Further Works

This paper describes early result of developing a procedure for dynamic selection of the ensemble members. The proposed procedure based on ensemble error prediction evaluated on water level prediction task using ensembles build from different combinations of three forecast sources. The mean RMSE for selected ensemble doesnt decrease significantly (8.3 cm in comparison to 9.0 cm for full ensemble, or only 2% in RMSE-SS score), which is result of relatively good forecasts, given by most of the ensembles (e.g. average RMSE for best available ensemble is 7.6 cm which gives only 5% increasing in RMSE-SS score). Nevertheless, the applied procedure enables to get the same or better result in 67% of the forecasts (including 56% of the strictly better forecasts) with the maximum obtained decrease of RMSE of 5.2 cm. Or, if we suppose that difference in errors in 0.1 cm is insignificant, we can get not worse results for 76% of the forecasts. This result shows usability of the proposed method but not a big difference between skill-scores should be improved in future studies.

Future work includes three main directions: a) more detailed study of selection procedure including study of different machine learning methods as decision trees and artificial neural networks; b) using different inputs for selection procedure that may describe forecasting system state; c) using different distance metrics to measure forecast error. In addition, a larger number

of models and other methods of ensemble calibration should be tested in experimental research.

Acknowledgements: This paper is supported by Russian Scientific Foundation, grant #14-11-00823.

References

[1] R Polikar. Ensemble based systems in decision making. *Circuits and Systems Magazine*, 2006.

[2] T.N. Krishnamurti, C. M. Kishtawal, Zhan Zhang, Timothy E. LaRow, David R. Bachiochi, C. Eric Williford, Sulochana Gadgil, and Sajani Surendran. Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, 13(23):4196–4216, 2000.

[3] Adrian E. Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using Bayesian Model Averaging to Calibrate Forecast Ensembles, 2005.

[4] Sergey V. Kovalchuk and Alexander V. Boukhanovsky. Towards Ensemble Simulation of Complex Systems. *Procedia Computer Science*, 51:532–541, 2015.

[5] Albert H R Ko, Robert Sabourin, and Alceu Souza Britto. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, 41(5):1735–1748, 2008.

[6] Tomasz Woloszynski and Marek Kurzynski. A probabilistic model of classifier competence for dynamic ensemble selection. *Pattern Recognition*, 44(10-11):2656–2668, 2011.

[7] HR Stanski, LJ Wilson, and WR Burrows. *Survey of common verification methods in meteorology*. 1989.

[8] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[10] Roman Schefzik, Thordis L. Thorarinsdottir, and Tilmann Gneiting. Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling. *Statistical Science*, 28(4):616–640, 2013.

[11] Mark S. Roulston and Leonard A. Smith. Evaluating Probabilistic Forecasts Using Information Theory, 2002.

[12] Jeffrey S. Whitaker and Andrew F. Loughe. The Relationship between Ensemble Spread and Ensemble Mean Skill, 1998.

[13] H. L. Cloke and F. Pappenberger. Ensemble flood forecasting: A review, 2009.