FORECASTING OF PROCESSES IN COMPLEX SYSTEMS FOR REAL-WORLD PROBLEMS

*Emil Pelikán**

tutorial

Abstract: This tutorial is based on modification of the professor nomination lecture presented two years ago in front of the Scientific Council of the Czech Technical University in Prague [16].

It is devoted to the techniques for the models developing suitable for processes forecasting in complex systems. Because of the high sensitivity of the processes to the initial conditions and, consequently, due to our limited possibilities to forecast the processes for the long-term horizon, the attention is focused on the techniques leading to practical applications of the short term prediction models. The aim of this tutorial paper is to bring attention to possible difficulties which designers of the predicting models and their users meet and which have to be solved during the prediction model developing, validation, testing, and applications. The presented overview is not complete, it only reflects the author's experience with developing of the prediction models for practical tasks solving in banking, meteorology, air pollution and energy sector.

The paper is completed by an example of the global solar radiation prediction which forms an important input for the electrical energy production forecast from renewable sources. The global solar radiation forecasting is based on numerical weather prediction models. The time-lagged ensemble technique for uncertainty quantification is demonstrated on a simple example.

Key words: *Forecasting, complex systems, data assimilation, multimodel forecasting, ensemble forecasting, judgmental forecasting, global solar radiation, pollution*

Received: November 11, 2014
Revised and accepted: December 12, 2014

DOI: 10.14311/NNW.2014.24.032

*Emil Pelikán, Faculty of Transportation Sciences, Konviktská 20, Prague 110 00, Czech Republic; Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, E-mail: pelikanemil@gmail.com

1. Introduction

Great effort has been devoted to the study of processes in complex systems on both the theoretical and applied sides during the last two or three decades. Complex systems represent a relatively new and widely interdisciplinary field passing through the physical, technical, environmental and socio-economic sciences. Growth rate of this field is rather fast, mainly because of progress in nonlinear modeling, statistical physics and computer science fields. There is no precise definition of what is meant by a complex system but most of the experts understand under the term complex system such a system that consists of many mutually interacting components (parts), and that shows emergent behavior, i.e., the collective behavior evinces some traits that cannot be easily derived or explained based on behavior of individual parts. Classic examples of complex systems can be found in economics (modeling of financial markets development), environmental science (the effects of environmental pollution), climatology (models of global warming), meteorology (weather forecasting), biology (models for spread of infectious diseases), transportation (traffic flow models), sociology (population growth) and in many other areas.

Study of complex systems behavior is difficult, and to understand the processes involved in them is often very hard. This is, among other things, due to the complexity of relationships and not infrequently even because of the obscurity of cause and effect relation (see the study of behavior of macroeconomic variables such as GDP, unemployment, inflation, etc.). Another reason is the interdisciplinarity for the full understanding of the patterns, knowledge of facts from several scientific fields (e.g., social science, computer science, mathematical statistics, physics, etc.) is needed. In spite of the diversity of different scientific disciplines, it turns out that complex systems show common universal features such as, e.g., nonlinear behavior, learning ability, etc.

Modeling of processes in complex systems can be divided into two basic approaches. The first involves the design and study of basic mathematical (and oftentimes simplified, but not simple) models via use of which it is possible to abstract the most important qualitative features of the complex systems behavior. Knowledge of the dynamical systems theory, networks, evolutionary computation, numerical methods, etc. is utilized.

The second approach is to develop behavioral models from measured data with use of methods of mathematical statistics, statistical learning, data mining, etc. Both approaches can surely be combined.

Knowledge of behavioral models will thereby allow us performing computer simulations, creating the scenarios of future development, forecasting the behavior for the short or medium term horizon, and using forecasts or obtained scenarios for management, planning and decision making.

In this paper, the attention is focused on the techniques important for development of high-quality models suitable for complex systems behavior forecasting especially on short term prediction horizon. Overview is not by a long shot complete, partly due to the limited range of this paper and partly due to the both still intensive development of new, and improvement of existing techniques and procedures. Presented techniques are mainly selected by virtue of the author's ex-

perience in real development of prediction models and their practical application in the fields of banking, meteorology and air quality, and energy industry.

2. Difficulty of the Prediction

In the past, the scientific thinking has been influenced by Newtonian paradigm in which the world can be reduced to a few basic elementary principles, allowing easy modeling and thus predicting the behavior through simple procedures. However, the study of behavior in microscopic and macroscopic worlds and other phenomena have shown us that the world is complex. Complexity of naturally occurring systems is a natural part of the world that surrounds us [13]. We can predict planetary motion for many thousands of years in advance, but we cannot reliably predict the evolution of GDP for the next year, the weather for several days ahead, development of share prices and and exchange rates for a few hours or minutes in advance.

Inability to accurately (or rather with sufficient and practically useful accuracy) predict the behavior of complex systems is mainly due to the incomplete information about the system, large number of parameters and our lack of ability to establish them, large number of variables and the inability to measure them, and last but not least the uncertainties contained in the executed measurements (measurement error).

The essential problem of complex systems processes prediction is the use of a priori information contained in partial (available) observation, and known (natural, technical, sociological, and other) laws. These laws can be expressed as a functional relationship between the various considered variables, and the task is to find values of the parameters so that the model predictions correspond to the measured variables as much as possible.

If we start from a deterministic description of the phenomena, which is based on the dynamical systems theory, studied processes are defined as vectors of time-dependent variables (states)

$$\boldsymbol{x}_t = (x_1(t), x_2(t), \ldots, x_n(t)), \tag{1}$$

where the dynamics of time evolution of the states is described by the equation

$$\boldsymbol{x}_t = f^t(\boldsymbol{x}_0, \boldsymbol{\lambda}) \tag{2}$$

and where $\boldsymbol{x}_t$ is state of the system in time t, $\boldsymbol{x}_0$ is initial state of the system, f^t is a smooth function describing the evolution of the state vector in time and $\boldsymbol{\lambda}$ is vector of (control) parameters. Thereby the states of the system may depend both on time and other variables (spatial dependence). Similarly, the vector of control parameters may depend on the time and possibly on other variables. For simplicity, however, we usually assume that the vector of parameters $\boldsymbol{\lambda}$ does not depend on time and that the dynamics of the process (2) may also be expressed by a system of partial differential equations

$$\frac{\partial x_i(t)}{\partial t} = f_i(\boldsymbol{x}_t, \boldsymbol{\lambda}) \tag{3}$$

where the function f_i is also time independent. If we know the f_i, and if we can somehow determine the value of the control parameters $\boldsymbol{\lambda}$, the question is why the prediction of time evolution of such a deterministic system should pose a problem.

For answering this question, there is in literature, see for example [2], [7], a nice example of very easy (so called logistic) model describing population growth given by equation

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = \alpha x_t \left(1 - \frac{x_t}{K}\right), \tag{4}$$

where x_t is the population size in a defined, limited space, α is a process parameter and K is the saturation coefficient. In the logistic model, a priori information, about the fact that the change of population size is proportional to the current population size, and that there is a saturation limit caused by limited space, is incorporated.

For discrete time, the following equation can be considered

$$x_{t+1} = \alpha x_t \left(1 - \frac{x_t}{K}\right). \tag{5}$$

In contrast to the simplicity of logistic model, system behavior described by equation (5) is not simple at all. Fig. 1 shows the course of time series generated by the logistic model for the first 100 values with the initial condition $x_0 = 0{,}2$ and with parameters $K = 1$ and $\alpha = 3{,}8$.

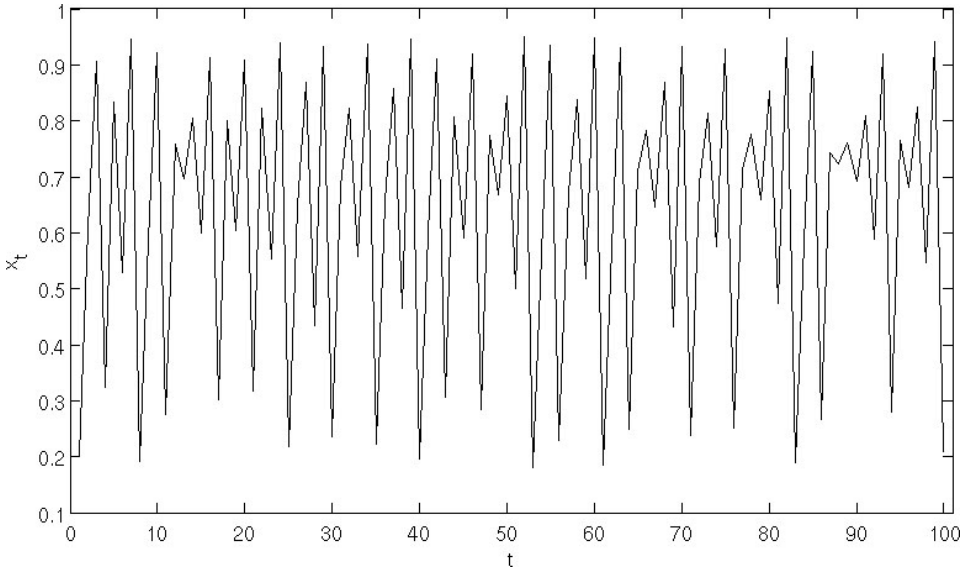


Fig. 1 *The course of time series (the first 100 values) generated by the logistic function with the initial condition* $x_0 = 0.2$.

It can be shown [2] that for a broad class of values of parameter α system (5) evinces chaotic behavior (Lyapunov exponent of the system is positive). That means, among other things, that the behavior of the system will be very sensitive to the value of initial conditions x_0. This fact is demonstrated in Fig. 2 where the

course of mean squared error of system (5) behavior prediction for different initial conditions in dependence on the prediction horizon is given

$$\mathrm{MSE}(m) = \frac{1}{N}\sum_{i=1}^{N}[F^m(x_0) - F^m(x_0 + \varepsilon_i)]^2, \tag{6}$$

where $F^m(x_0)$ is evolution of the system (5) from the initial conditions x_0, m steps ahead and ε_i are randomly selected values of uniform distribution on the interval $(-0.0001; +0,0001)$ respectively $(-0.001; +0,001)$. For the calculation, 10 000 realizations of the random error for each initial condition x_0 was used, whereas the initial conditions took different values from 0.005 to 0.995, and the resulting mean squared error was averaged.

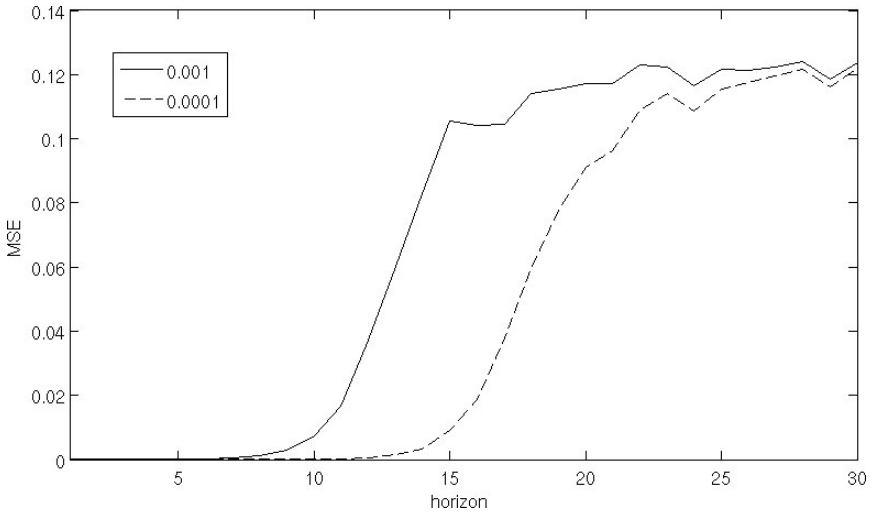


Fig. 2 *The course of mean squared error MSE in dependence on prediction horizon m for logistic function $x_{t+1} = 3.8\ x_t(1 - x_t)$ with randomly noised initial conditions x_0.*

In Fig. 2 can be clearly seen the exponential growth of prediction error with increasing predicting step (up to the saturation limit). The experiment demonstrates the difficulty of behavior predictability of chaotic complex systems for a longer prediction horizon, without knowing the initial state precisely enough, even assuming that the dynamics of the system is perfectly known, including the values of the parameters involved. Besides, the effect of inaccuracy in the parameters estimation can also cause exponential growth of prediction error, as it is demonstrated in Fig. 3.

The problem of the inability to measure the state of the system with sufficient accuracy is in some cases complicated by the fact that by virtue of some technical or principal reasons we cannot measure most of the state variables, and the measurement is performed only on a very limited part of the state variables. From the dynamical systems theory results that the system dynamics can be to a certain

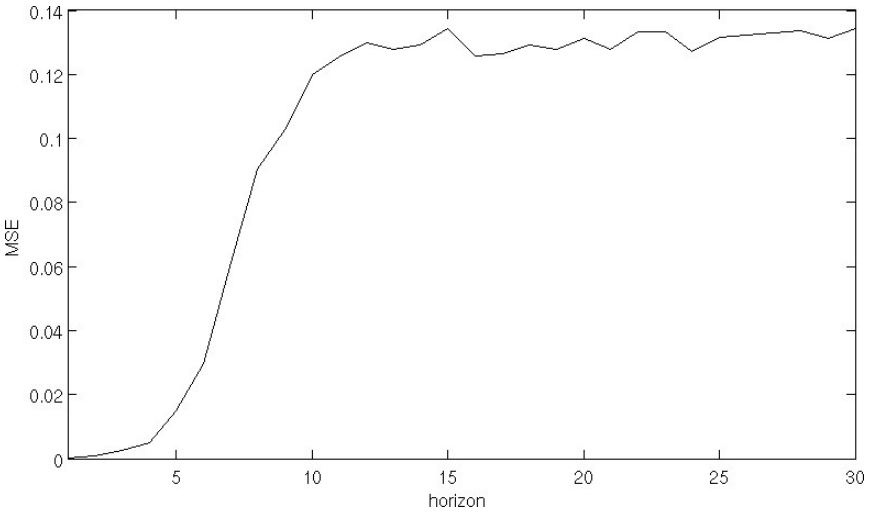


Fig. 3 *The course of mean squared error MSE for logistic function* $x_{t+1} = \lambda x_t(1 - x_t)$ *with different values of parameter* λ *randomly generatedy from interval (3.7; 3.9) in dependence on prediction horizon* m.

extent examined solely on the basis of the single state dynamics. If we assume a deterministic system with dimension of n states, see the Eq. (1).

Takens' theorem [21] implies that the dynamics of such system (attractor) can be reconstructed from vectors of time-shifted (shift T) states of single variable with nested dimension d_E

$$\boldsymbol{x}_{t,1} = (x_1(t), x_1(t+T), \ldots, x_1(t+[d_E-1]\,T)), \tag{7}$$

where $d_E > 2n+1$. Of course if n is very high, the nested dimension is even larger. Yet in many applications the dynamics of the whole system via nested vectors with much lower dimension can be examined [7].

Predicting processes in systems with chaotic behavior is therefore difficult, but not impossible. It is just necessary to respect some limitations, particularly the high sensitivity to the initial conditions and the resulting inability to achieve low prediction errors for a longer time horizon, as well as generally the need of non-linear models, and complex procedures and techniques leading to the successful prediction.

3. General Prediction Model

As it was outlined in the previous section, the essential problem of complex systems behavior prediction is the inability to measure the states of the system with sufficient accuracy. A common objective of the prediction methods development is to predict the time evolution of the complex system states usually for the short term horizon (see problems of forecasting the systems with chaotic behavior for a longer time horizon as described in the previous section) under the real assumption that

we do not have or cannot use the complete information about system dynamics, the parameters must be estimated from the measured data, and furthermore accessible measurements are subject to measurement error. Moreover, we often face the situation in which we are not at all able to measure, resp. observe, some states of the system, and sometimes it is even a large majority of all states (see the next section dealing with data assimilation). If we limit ourselves to discrete systems and prediction horizon one step forward, the task of prediction can be simplified to finding the prediction function (prediction model) F^{f}

$$\boldsymbol{x}_t^{\mathrm{f}} = F^{\mathrm{f}}(\boldsymbol{x}_{t-}, \boldsymbol{\theta}) \tag{8}$$

where $\boldsymbol{x}_{t-}$ is history of (measured) states

$$\boldsymbol{x}_{t-} = (\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \ldots, \boldsymbol{x}_0) \tag{9}$$

and $\boldsymbol{\theta}$ is vector of parameters. If we need to predict k steps ahead ($k > 1$), prediction model F^{f} can be used in following way:

$$\boldsymbol{x}_{t+k-1}^{\mathrm{f}} = F^{\mathrm{f}}(\boldsymbol{x}_{t+k-2}^{\mathrm{f}}, \boldsymbol{x}_{t+k-3}^{\mathrm{f}}, \ldots, \boldsymbol{x}_t^{\mathrm{f}}, \boldsymbol{x}_{t-}, \boldsymbol{\theta}). \tag{10}$$

In real practice, we often limit ourselves to prediction models that do not use the whole history of the measured states, but only a certain part of the actual history, for example

$$\boldsymbol{x}_t^{\mathrm{f}} = F^{\mathrm{f}}(\boldsymbol{x}_{t-,h}, \boldsymbol{\theta}), \tag{11}$$

where

$$\boldsymbol{x}_{t-,h} = (\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \ldots, \boldsymbol{x}_{t-h}). \tag{12}$$

Development of specific form of prediction model F^{f} can be based on several principles resulting from for example Box-Jenkins methodology, principles of artificial neural networks, fuzzy approaches, Bayesian and SVM (support vector machines) techniques and many others. This extensive and interesting topic, however, is not the aim of this paper and therefore those who are interested can refer to the rich academic literature, e.g., [1], [3], [11], [17], [20].

4. Data Assimilation Technique

No matter what the general prediction model (described in the previous section) will look like, it is necessary to ensure the accessibility of input values, i.e., the states $\boldsymbol{x}_{t-}$ needed for the calculation of prediction $\boldsymbol{x}_t^{\mathrm{f}}$. As aforementioned, in some systems, due to some technical or principal reasons, we cannot measure most of the state variables, and the measurement is performed only on a very limited part of the state variables. If we need to reconstruct all the states of the system from the available measurements (that naturally contains noise), we can use the technique of data assimilation. It is based on a modification of the initial (a priori) estimation of the states according to the error of the observing operator and available measurements (see, e.g., [9]). Assuming that the initial estimation of the states in the system is $\boldsymbol{x}_t^{\mathrm{p}}$, $\boldsymbol{H}$ is the observing operator transforming (or interpolating) the

system states $\boldsymbol{x}_t^{\mathrm{p}}$ to set of observed states $\boldsymbol{x}_t^{\mathrm{o}}$, then the newly renovated state of the system $\boldsymbol{x}_t^{\mathrm{new}}$ via using measurements $\boldsymbol{x}_t^{\mathrm{o}}$ is given by relation

$$\boldsymbol{x}_t^{\mathrm{new}} = \boldsymbol{x}_t^{p} + \boldsymbol{W}\left[\boldsymbol{x}_t^{o} - \boldsymbol{H}(\boldsymbol{x}_t^{p})\right], \tag{13}$$

Where the matrix $\boldsymbol{W}$ is weighing matrix, which is usually determined on the basis of covariance estimates of statistical errors of observation and system states. For the initial estimate of system states may be used the variant of prediction model (8), therefore we consider

$$\boldsymbol{x}_t^{\mathrm{p}} = \boldsymbol{x}_t^{\mathrm{f}}. \tag{14}$$

There are many methods of data assimilation based on equation (13), such as methods on the basis of the optimal interpolation (OI) technique, the 3D-Var, 4D-var [4], [9].

As an example, it is mentioned one of the most frequently used methods, the method utilizing the technique of Kalman filters. We assume that the system states $\boldsymbol{x}_t$ are random, and that the corresponding random process has the Markov property, i.e., it is met the equality

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-,h}) = p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}). \tag{15}$$

Further assume that the reconstruction of states $\boldsymbol{x}_{t-1}^{\mathrm{new}}$ is an unbiased estimation, i.e., it holds

$$\boldsymbol{x}_{t-1}^{\mathrm{new}} = \boldsymbol{x}_{t-1} + \boldsymbol{\eta}_{t-1}, \tag{16}$$

where the mean value of the random error in $E(\boldsymbol{\eta}_{t-1}) = \boldsymbol{0}$, where E is the operator of the mean. We assume further that the prediction model used for the prediction of the initial state of the system and the observing operator $\boldsymbol{H}$ are linear, i.e.,

$$\boldsymbol{x}_t^{\mathrm{p}} = \boldsymbol{M}_t\, \boldsymbol{x}_{t-1}^{\mathrm{new}} \tag{17}$$

and

$$\boldsymbol{H}(\boldsymbol{x}_t^{\mathrm{p}}) = \boldsymbol{H}_t\, \boldsymbol{x}_t^{\mathrm{p}}, \tag{18}$$

where $\boldsymbol{M}_t$, $\boldsymbol{H}_t$ are the matrix of parameters. Label then

$$\boldsymbol{\varepsilon}_t = \boldsymbol{x}_t^{\mathrm{o}} - \boldsymbol{H}_t\, x_t, \tag{19}$$

$$\boldsymbol{\nu}_t = \boldsymbol{x}_t - \boldsymbol{M}_t\, x_{t-1}. \tag{20}$$

Covariance matrices will be labeled

$$\mathrm{cov}(\boldsymbol{\eta}_{t-1}) = \boldsymbol{P}_{t-1}^{\mathrm{new}}, \tag{21}$$

$$\mathrm{cov}(\boldsymbol{\nu}_t) = \boldsymbol{Q}_t, \tag{22}$$

$$\mathrm{cov}(\boldsymbol{\varepsilon}_t) = \boldsymbol{R}_t. \tag{23}$$

Assuming that $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\nu}_t$ are uncorrelated and also have a zero mean value, following recursive relations can be derived

$$\boldsymbol{x}_t^{\mathrm{new}} = \boldsymbol{x}_t^{\mathrm{p}} + \boldsymbol{K}_t\left(\boldsymbol{x}_t^{\mathrm{o}} - \boldsymbol{H}_t\, \boldsymbol{x}_t^{\mathrm{p}}\right), \tag{24}$$

$$\boldsymbol{P}_t^{\mathrm{new}} = (\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{H}_t)\boldsymbol{P}_t^{\mathrm{p}}, \tag{25}$$

$$\boldsymbol{P}_t^{\mathrm{p}} = \boldsymbol{M}_t\, \boldsymbol{P}_{t-1}^{\mathrm{new}}\, \boldsymbol{M}_t^{T} + \boldsymbol{Q}_t, \tag{26}$$

where

$$\boldsymbol{K}_t = \boldsymbol{P}_t^p \boldsymbol{H}_t^T \left(\boldsymbol{H}_t \boldsymbol{P}_t^{\mathrm{p}} \boldsymbol{H}_t^T + \boldsymbol{R} \right)^{-1}. \tag{27}$$

These relations represent the classical equations of the Kalman filter, whereas the matrix $\boldsymbol{K}_t$ is called Kalman gain.

In light of the complex system processes prediction model development, it is good to notice, that the prediction model (17) is linear, but parameter matrix can time dependent (so called adaptive model).

If the all considered distributions are normal, the resulting distribution is

$$p(\boldsymbol{x}_t^{\mathrm{new}} | \boldsymbol{x}_t^{\mathrm{o}}) \tag{28}$$

also normal with meanvalue $\boldsymbol{x}_t^{\mathrm{new}}$ and covariance matrix $\boldsymbol{P}_t^{\mathrm{new}}$. Under the aforementioned assumptions, not only the "mean value" mean value of the new state, but also the entire probability distribution (see Fig. 4) and the levels of reliability (quantile forecast) derived from it, can be relatively easily predicted. Kalman fil-

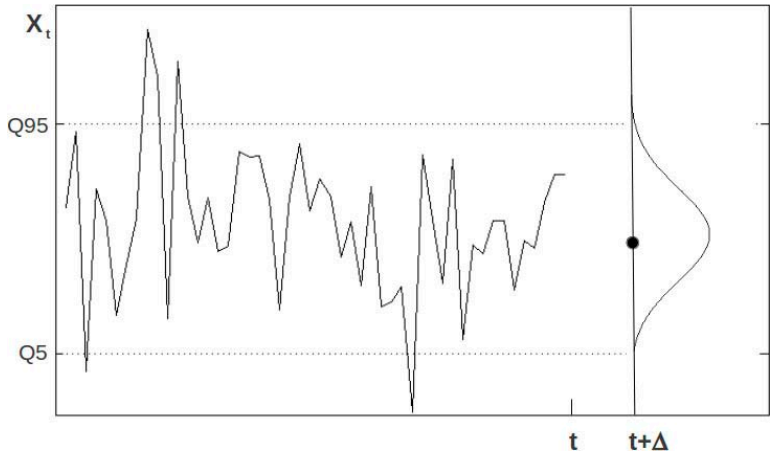


Fig. 4 *Point, probability forecasting and prediction 5% a 95% quantiles – illustrative figure.*

ter's equations (17)–(27) also allow prediction of measured states $\boldsymbol{x}_t^{\mathrm{o}}$ and are the foundation of prediction modeling technique called state-space, see, e.g., [3], [17].

5. Optimal Prediction Model

Typically, in complex systems, for the same task there can be developed several prediction models usable for the real forecast. As in many scientific areas, the best possible solution (optimal model) usually cannot be found; therefore prediction model development is an inventive process. Following quotation attributed to G. Box and W. E. Deming [14] lends itself to the situation very well: "Essentially, all models are wrong, but some are useful."

If a suitable objective function, that evaluates the prediction quality of our model, is given by discrepancy between predicted and actual values, the best model

can be assessed according to the values of the objective function. The basic problem is that the value of this function can be calculated only on the basis of known facts (based on historical data). To believe that our model will be the best in the future – and that matters – we have to make certain assumptions about our system and the data that it generates.

If we study prediction model in the shape $\boldsymbol{x}_t^{\mathrm{f}} = \boldsymbol{F}^{\mathrm{f}}(\boldsymbol{x}_{t-,h}, \boldsymbol{\theta})$, prediction function development and estimation of its parameters are the fundamental tasks. Prediction function structure may be determined a priori (based on a priori knowledge of the system behavior) or from data, e.g., a suitable choice of class of sufficiently universal functions, as is the case of for example usage of artificial neural networks. Parameters $\boldsymbol{\theta}$ then can be determined based on minimization of suitably chosen objective resp. loss-function. For concrete measured values of state variables $\boldsymbol{x}_t$, $t = 1, 2, \ldots, N$, loss-function be considered

$$L(\boldsymbol{\theta}) = \left\| \boldsymbol{x}_t - \boldsymbol{F}^{\mathrm{f}}(\boldsymbol{x}_{t-,h}, \boldsymbol{\theta}) \right\|^2 . \tag{29}$$

Since we are looking for such a parameter $\boldsymbol{\theta}$ for which (29) is minimal, we can write

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}). \tag{30}$$

This procedure is called empirical risk minimization – ERM. Because of the prediction function $\boldsymbol{F}^{\mathrm{f}}$ can be very complicated (nonlinear etc.), looking for a suitable $\boldsymbol{\theta}$ minimizing $L(\boldsymbol{\theta})$ can be nontrivial task. The essential problem of ERM procedure is that parameter $\boldsymbol{\theta}$ estimation on the basis of available data $\boldsymbol{x}_t$, $t = 1, 2, \ldots, N$, probably will not be identical with estimation made on data measured during next period, even assuming that the function $\boldsymbol{F}^{\mathrm{f}}$ does not change in time. It can happen that the value of loss-function $L(\boldsymbol{\theta})$ in the next period increases that much that model becomes inutile. Thereby we concern not only over the value $L(\boldsymbol{\theta})$ on given data, which is cold training error (or in-sample error), but also over the information how the loss-function will evolve on data that are going to be obtained. This fact can be expressed that we care about "expected value" of loss-function $E[L(\boldsymbol{\theta})]$, where E is mean value operator. Function $E[L(\boldsymbol{\theta})]$ is called generalization error (or out-of sample error). The reason of training error and generalization error difference is fact that training data $\boldsymbol{x}_t$, $t = 1, 2, \ldots, N$ are not representative for behavior of the studied system (but are just some kind of a behavior demonstration determined by the selected data obtained with technical, physical or other limitations) and furthermore the available data (measurements) are not perfect (are noised).

Problem of overfitting the model can be to a certain extend removed by using the model validation techniques (out-of sample testing) and regularization of minimization process.

5.1 Model Validation

Within the model validation we estimate generalization error with using the second if possible independent sample which does not enter into the minimization process (30). Thus normally we divide the available data into training and validation groups on which we assess the loss-function values. To split the data into training

and validation can be done by different ways (randomly, sequential selection of some part of the data, etc.).

If we use numerical iterative procedures in ERM procedure (minimization of the loss function), a course of empirical error and estimation of generalization error can be seen during the iterative process. While empirical error usually decreases (depending on the chosen numerical procedure), the estimation of loss function generalization error calculated on a validation sample after a certain time reverses the trend (from decreasing to increasing), as the Fig. 5 illustrates. At this point it is appropriate to stop the iterative process and use as a parameter estimation that $\boldsymbol{\theta}$, that minimizes the course of generalization error estimate (early stopping criterion).

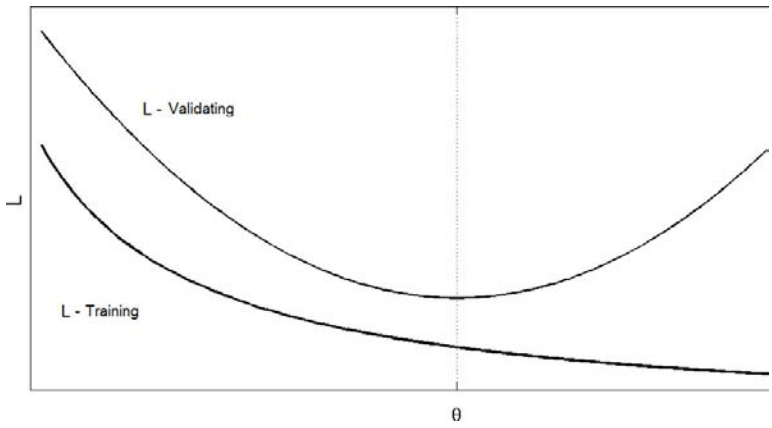


Fig. 5 *Parameter $\boldsymbol{\theta}$ estimation with using loss-function course on training and validating data (early-stopping criterion).*

5.2 Regularization

The problem of minimizing $L(\boldsymbol{\theta})$ can be, and usually indeed it is, ill-conditioned, i.e., with minor changes in the input data we can receive large changes in the estimations of parameter $\boldsymbol{\theta}$. That is why the minimization is

$$L(\boldsymbol{\theta}) + \lambda\, d(\boldsymbol{\theta}), \tag{31}$$

where the function $d(\boldsymbol{\theta})$ is regularization or penalty function.

Via the appropriate choice of regularization (penalty) function, we can reduce the complexity of the model and thereby reduce the models generalization error. This principle is in line with intuitive understanding that complex (over-parametrized) models may more nearly reflect the characteristics of learning sample data (including noise in them), and not the general properties of the process that needs to be predicted. At the same time, this principle is in line with philosophical concept known as Occam's razor principle, which says that "entities must not be multiplied beyond necessity", therefore that from the models with acceptable predictive performance on training data we select the easiest one.

Example of penalty function is function

$$K\frac{C(N)}{N}, \tag{32}$$

where K is number of model parameters and N is range of data. For $C(N) = 2$ we get the often used Akaike information criterion AIC, [17].

5.3 Testing the Model

If the model is built with using the training and validation data, it is recommended to independently test it on the test data set, which does not contain the data from the training nor the validation set. Objective (loss) function for testing the model can take many forms reflecting user requirements for quality of prediction models, the fulfillment of which may not be trivial. It is also good to realize that in many practical applications the objective function for testing the models may not be the same objective function for model development in the phase of training and validation.

For simplicity, we assume here that our goal is the prediction of only one component (state variable) of a complex system and that we test the prediction model on the test set x_t, $t = 1, 2, \ldots, N$.

Mean squared error MSE is then given by formula

$$L_{\mathrm{MSE}} = \frac{1}{N}\sum_{t=1}^{N}\left(x_t - x_t^{\mathrm{f}}\right)^2, \tag{33}$$

where x_t^{f} is forecast of variable x_t. To compare the quality of models by MSE has many advantages because of that minimization L_{MSE} (so-called method of least squares) is widely used to prediction model development. Generalization of the objective function (33) is the Minkowski objective function, which is given by

$$L_{\mathrm{M}} = \frac{1}{N}\sum_{t=1}^{N}\left|x_t - x_t^{\mathrm{f}}\right|^R. \tag{34}$$

It is obvious the for $R = 2$ the function (34) is identical with function (33). For $R = 1$ we get the mean absolute error MAE.

In order to compare the predictive performance of models for forecasting the variables with different mean value (e.g. for forecasting macroeconomic variables for differently sized regions), very popularly used mean absolute percentage error MAPE is defined as:

$$L_{\mathrm{MAPE}} = \frac{100}{N}\sum_{t=1}^{N}\left|\frac{x_t - x_t^{\mathrm{f}}}{x_t}\right|. \tag{35}$$

In the L_{MAPE} denominator an absolute value of the predicted variable occurs that at very low values close to zero leads to very high values of this criterion, therefore, it is necessary to evaluate the percentage error with great care.

Evaluation of the model quality according to large deviations is another example of objective function expressed by formula

$$L_{\mathrm{MAX}} = \frac{1}{M} \sum_{|x_t - x_t^{\mathrm{f}}| > \varepsilon} \left| x_t - x_t^{\mathrm{f}} \right|, \tag{36}$$

where ε is preselected acceptable value of error and M is the number of prediction errors larger (in absolute value) than ε.

In the literature about predictions [11] furthermore the criterion appears that compares model performance with so called persistent model

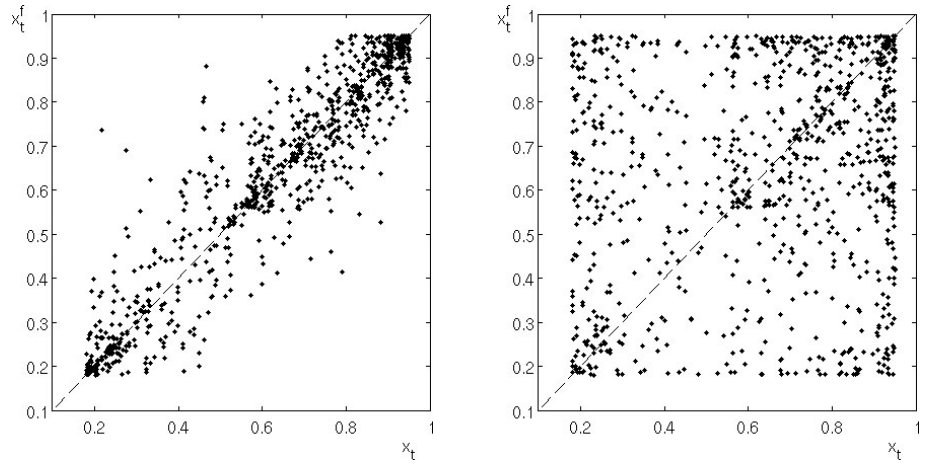
$$x_t^{\mathrm{f}} = x_{t-1}. \tag{37}$$

From the equation (37) implies that persistent models predict the closes future by last known state of the system. The objective function (so called Theil's U-statistics) is then defined as

$$L_{\mathrm{U}} = \sqrt{\frac{\sum_{t=1}^{N} \left(\frac{x_{t+1} - x_{t+1}^{\mathrm{f}}}{x_t} \right)^2}{\sum_{t=1}^{N} \left(\frac{x_{t+1} - x_t}{x_t} \right)^2}}. \tag{38}$$

For models with better performance than the persistent models is then worth $L_{\mathrm{U}} < 1$. In a similar way, we can compare the performance of the prediction model with a model type like random walk, etc.

As a final example of evaluating the quality of a model, use the cross diagrams which are shown in the graphs in Fig. 6



(a) (b)

Fig. 6 *Cross diagram of predicted vs. actual value for logistic function* $x_{t+1} = 3.8\, x_t(1 - x_t)$ *with randomly noised initial condition* x_0 *for prediction step* $m = 10$ *(a) and* $m = 15$ *(b).*

6. Multi-model Prediction

Complexity of processes prediction in complex systems is caused by that multiple models (based on the same or different principles) can be developed. The models can be tested on learning and validation data and then choose the best one from the portfolio. However, model development is difficult process, so the question why not to use the properties of other (already existing) models arises. The principle of multi-model prediction is based on that; when the predictions from different models are combined, whereas it often happens that the final prediction has better performance (lower prediction error) than the prediction performance of each model individually.

Assuming that multi-model prediction is made up of the K prediction arithmetic average of K prediction models

$$x_t^{\mathrm{MIX}} = \frac{1}{K}\sum_{j=1}^{K} x_{t,j}^{\mathrm{f}}, \tag{39}$$

then mean squared error of combined forecast is equal to

$$E\big(x_t^{\mathrm{MIX}} - x_t\big)^2 = E\left(\frac{1}{K}\sum_{j=1}^{K} x_{t,j}^{\mathrm{f}} - x_t\right)^2 = E\left(\frac{1}{K}\sum_{j=1}^{K} \varepsilon_{t,j}\right)^2, \tag{40}$$

where $\varepsilon_{t,j}$ is prediction error of j-th prediction model. Assuming that errors of the individual models are uncorrelated with zero mean value and with identical variance σ^2, then holds

$$E\big(x_t^{\mathrm{MIX}} - x_t\big)^2 = \frac{1}{K^2}\sum_{j=1}^{K} E(\varepsilon_{t,j}^2) = \frac{1}{K}\sigma^2, \tag{41}$$

where E is mean value operator. The variance of the resulting error process is this way reduced (under certain preconditions) by factor K. Generally, the errors of individual models are correlated, however, in many cases the multi-model prediction reduces the prediction errors.

Of course the formula (39) for combining can be generalized so that the final prediction is a weighted combination of individual predictions,

$$x_t^{\mathrm{MIX}} = \sum_{j=1}^{K} w_j x_{t,j}^{\mathrm{f}}, \tag{42}$$

where the weights are set so they minimize the value of the loss-function.

The aforementioned principle can be applied to an active development of individual models suitable for combined forecast, for example, that to a given portfolio of models another one is added while demanding that its prediction error is "decorrelated" with the error of current portfolio [15]. Schematically, this procedure is shown in Fig. 7. It is also possible to estimate the weights and individual model parameters simultaneously (see so called mixture of experts models [1]).

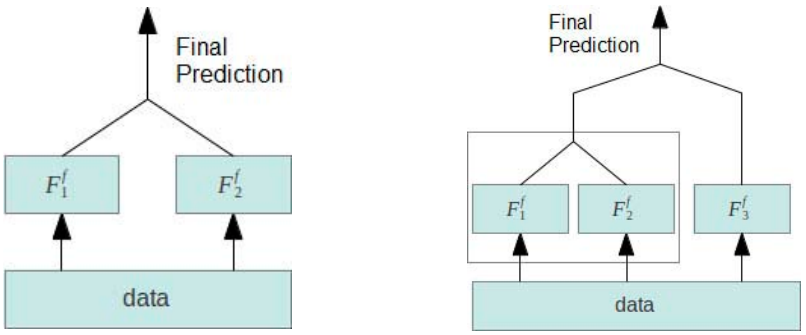


Fig. 7 *Scheme of classic combined forecast made by two prediction models (a) and adding new model to current portfolio (b).*

To the category of multi-model prediction we may include a technique called MOS (model output statistics), which is based on statistical processing (postprocessing) the outputs from the physical (biological, chemical and others) models. The predictions of these models (often very sophisticated) may be deviated, prediction errors are correlated etc. The aim of MOS technique is to develop a statistical model improving the statistical characteristics of the original models. Schematically, the use of MOS technology is shown in Fig. 8.

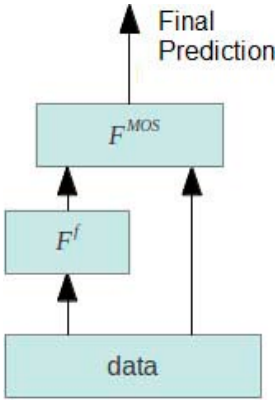


Fig. 8 *Statistical improvement of prediction performance of prediction models* $\boldsymbol{F}^{\mathrm{f}}$ *via MOS (model output statistics) technique.*

7. Ensemble Forecasting

As discussed in Section 2, the trait of complex systems is the high sensitivity of further development to the initial conditions. The essence of ensemble forecasting is acquiring more individual predictions (predicted trajectories) by suitable perturbations of the initial state of the system. This procedure is schematically shown in Fig. 9.

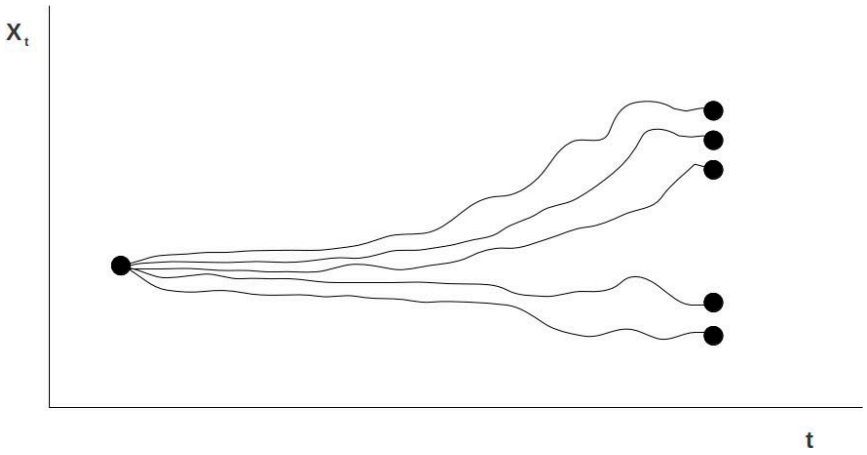


Fig. 9 *Acquiring the ensemble forecasts by perturbation of the initial state of the complex system.*

From these forecasts can be obtained the final forecast (for example as a weighted combination of individual forecasts as in the case of multi-model predictions described in the previous section) as well as the information on the uncertainty of given forecasts. Acquired ensemble forecasts should be, ideally, a random selection from the probability distribution, which corresponds to the evolution of given system. The best solution would be to describe the time evolution of the investigated system using the equations describing the evolution of the probability density $p(\boldsymbol{x}_t)$ as it is explored in the theory of stochastic processes and as it is indicated in Section 4., and to provide the forecast in the form of predicted probability density. However, in real situations it is often difficult to meet all necessary prerequisites and reliably estimate the theoretical probability density. Ensemble forecasting thus represent a useful step from "point forecasting" to fully probabilistic predictions. Besides, from practical point of view (according to the author's experience) it is not easy to communicate outputs of probabilistic forecasting models to a normal user in accessible way.

Another method is construction of so-called structure time-lagged ensemble, where the individual prediction ensemble for the time $t + k$ is created from predictions for the same time but obtained in the times t, $t + 1$ a $t + 2, \ldots, t + k - 1$ (see Fig. 10). The advantage of time-lagged ensembles is that we do not need to do any additional predictive calculations (which for some complicated models may represent significant time savings), because we usually predict to a sufficiently long time horizon and at each step the forecasts is updated. More detailed information about ensemble forecasting can be found e.g. in [9].

8. Judgment (Expert) Forecasting

The human brain despite advances in neuroscience and artificial intelligence is still the most perfect tool that can be used either on its own or in combination with

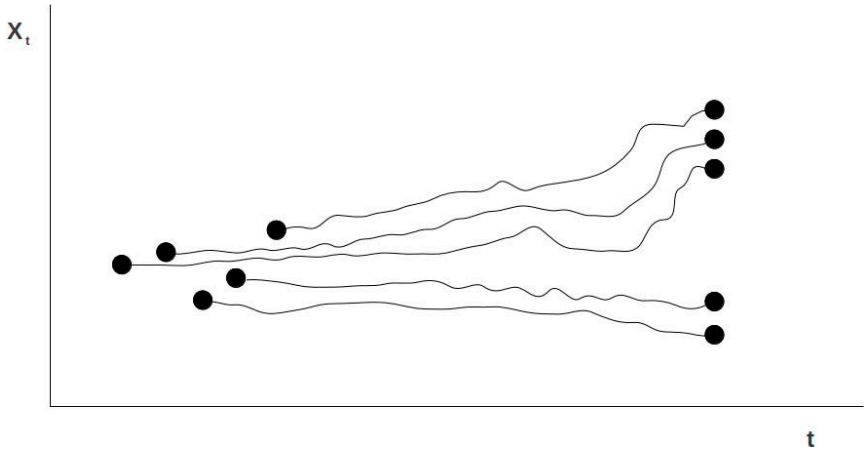


Fig. 10 *Acquiring ensemble forecasts via time-lagged predictions.*

supporting tools for its decision. As already has been indicated in the introduction, complex system behavior prediction is not an easy task. Even with usage of sophisticated nonlinear prediction methods there is a need to correct the predictions in an appropriate manner and correction made by human expert (experts) is one of the most efficient options. Of course, there are situations where judgmental prediction is not possible (for control of complex devices in real time, for prediction of many hundreds or thousands of variables behavior at one minute intervals, etc.).

The complex systems behavior prediction made by human experts is characterized by the fact that in the vast majority of cases there is not unambiguous conclusion (see for example the issue of global warming). It is due to the complexity of relations and even because of the obscurity of cause and effect relation (see the study of behavior of macroeconomic variables such as GDP, unemployment, inflation, etc.). Differences in forecasts are given by the different experiences of individual experts (each uses a different set of learning). Moreover, existing relations, if they have existed in the past, are changing in time (the world is changing), often very dramatically.

Interesting is also the tendency of thinking to linear extrapolation, while nonlinear relations between multiple variables are very difficult to understand (people like simple explanations such as simple implication). One of the most fundamental influences on judgmental predictions about the complex systems behavior is the problem of expectation of given evolution that expresses itself in subconscious filtering of the obtained data so that they meet those expectations. Even the results of a very sophisticated prediction models can be suppressed or interpreted in different ways so that predictions do not diverged from the expected framework. This could be for example one of the possible reasons of wrong forecasting the economic crisis in the last four years, see [6].

Descriptive illustration of expectation influence on the interpretation of image information is shown in Fig. 11.

Fig. 11 *Double interpretation of the same visual information – young girl vs. old woman (source internet).*

Another phenomenon that affects expert prediction is too high self-confidence of some experts or to small self-confidence and their easy suggestibility by voices of other experts.

Yet the expert predictions are very useful especially in combination with mathematical models, and that is because the expert is able to intuitively use for prediction such information whose usefulness may not be apparent at first sight and which are not included in the mathematical model (or it was not possible to include them due to economic or technical reasons). Furthermore, in certain areas the expert prediction indispensable (medicine) or is required (e.g. in expert opinions) by the users (e.g. public administration).

9. Experiment with Forecasting the Solar Power

Problem of weather forecasts problem is closely tied with the modeling of complex natural processes. These problems are of great importance for almost all transportation activities, maybe with excerption of subways (metros).

The chaotic nature of the system describing the atmospheric phenomena has been described in the classical work of E. Lorenz [10], which led to the birth of a new scientific discipline and the intense study of the deterministic chaos properties. Due to the sensitivity of weather to initial conditions the predictability of weather is limited. At present, it is considered that the theoretical limit of the forecast horizon for the weather forecasts is about 3–4 weeks [9]. With increasing computing power, refinement of initial conditions, increasing number and quality of observation, the use of modern statistical methods, methods for data assimilation and ensemble forecasting techniques, today a practical time of weather predictability (i.e. when the operational forecast is better than climatological) is about 15 days. However,

this does not apply to meteorological variables with regional character (like the solar power is), where the practical predictability horizon is much lower.

Solar power forecasting whether for the short term (1–72 hours) or long-term (using the climatological values) horizons in fine regional resolution is an important task in the context of a sharp increase of photo-voltaic installations all over the Europe and the Czech Republic. While the long-term forecasts are used for estimating the economic return on investment into the photo-voltaic farms, the short-term forecasts are matter of interest in companies engaged in the production and distribution of electrical energy, the energy trader, small operators of the photo-voltaic system and other.

Task of forecasting the solar power is not easy. Besides the local nature, it can be a problem even that there is a need to predict the values of this variable with a relatively fine time step (typically 1 hour but even less) and in fine regional resolution (below cca 10 km). Such a prediction is not possible without the use of numerical weather prediction (NWP) models. These are based on numerical solution of partial differential equations describing the physical processes in the atmosphere (flow dynamics, atmospheric water vapor flux, energy flows, etc.). Because of the very complicated behavior of the atmosphere, there is no universal best model of weather. Therefore there are different types of NWP models that differ in the spatial and time scale, level of details in physical phenomena description, idealized models of flow in landscape, and various approaches for modeling the cloud microphysics, etc. While describing the processes in the atmosphere in NWP models, the time evolution of the state vector comprising of spatially discretized values of physical quantities such are the pressure, temperature, wind speed and direction, humidity, concentrations of other substances in the air, etc., is described.The model describes the states evolution always in reference to specific area (domain), which is divided into tens of vertical layers and horizontally is covered mostly by square networks. The vertical division is in NWP models usually solved by changes the hydrostatic pressure (in the so called σ-layers). The time step for modeling the physical phenomena varies depending on the horizontal resolution about 10–90 seconds, user outputs are typically provided with a time step of 1 hour.

To predict the solar power, we used numerical prediction models implemented in the system MEDARD [5], [12], that is operationally run in the Institute of Computer Science, Academy of Sciences of the Czech Republic, from 2004. The core of the system is the NWP model WRF (Weather Research and Forecasting), which replaced the older version of the model MM5 (Mesoscale Model, version 5). To their development a number of organizations such as the National Center for Atmospheric Research (NCAR), National Oceanic and Atmospheric Administration (NOAA), National Centers for Environmental Prediction (NCEP) and others, have contributed. WRF and MM5 models are primarily focused on the evolution of regional weather in 2–20 km resolution, but experimentally are investigated their properties at a resolution of 1 km and a smaller [8]. The boundary conditions are derived from global atmospheric model GFS, which currently has a vertical resolution of about 50 km. They are exhibited in 6 hour intervals when actual measurement of states predicted in the previous step via using data assimilation techniques is used. Here we use information from ground-based weather stations, radar measurements, weather balloons and measurements on aircrafts and ships.

For simulations in fine resolution, nesting of several domains is calculated. There are three nested domains used in the system MEDARD, in resolution 27 km, 9 km and 3 km.

Outputs from the models MM5 and WRF were used as inputs to the prediction model for forecasting the electricity production from photo-voltaic farms in the Czech Republic operated by CEPS, a.s. Forecast was carried out on the forecasting horizon "D+1", thus predicts the production of electricity for the next day in the morning (between approximately 7–8 hours) of the current day.

Smoothed course of prediction error normalized to the installed power is shown in Fig. 12. Of course, we see that, paradoxically, the older version of NWP model MM5 shows better results than the newer version of the NWP model WRF. It is due to the fact that NWP models are not a priori developed for the specific (in our case energy) needs, but are developed to predict the weather as a whole. It also confirmed that multi-model prediction (in our case, the final prediction is calculated as the arithmetic mean of the outputs of both models and is shown in Fig. 12 marked as MIX) may improve the prediction.

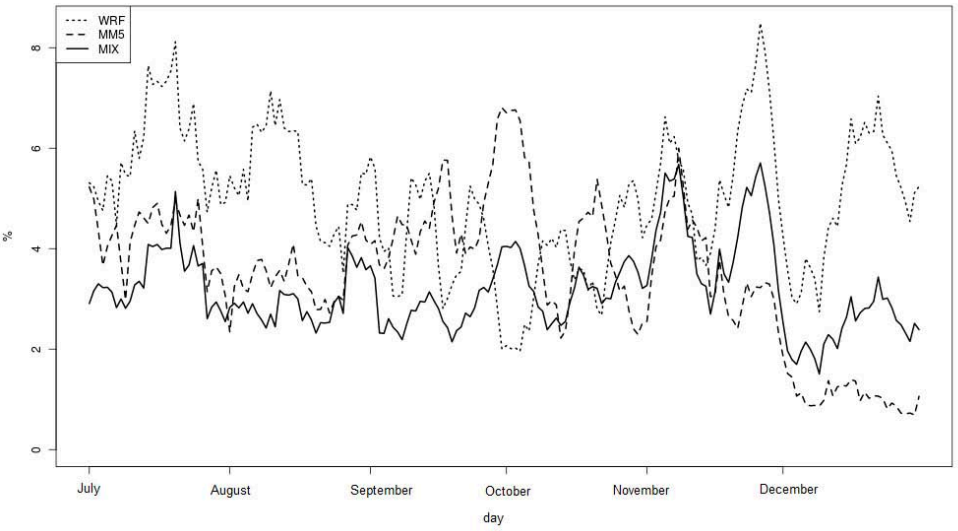


Fig. 12 *Course of prediction error of electricity production from photo-voltaic sources in the Czech Republic during the second half of year 2011 (error expressed in percent with respect to installed power) for predictive horizon D+1 (morning forecast for the next day) with use of numerical weather prediction models WRF and MM5 and their combination (MIX).*

In Figs. 13 and 14 an illustration of time-lagged ensemble forecast gradually developed from 14 predictions of NWP model MM5 in six hours step for two selected days can be seen. The predictions are compared with measurements of solar power on the weather station of AIM (automatic immission monitoring) network operated by the Czech Hydrometeorological Institute. Increased variability of predictions is

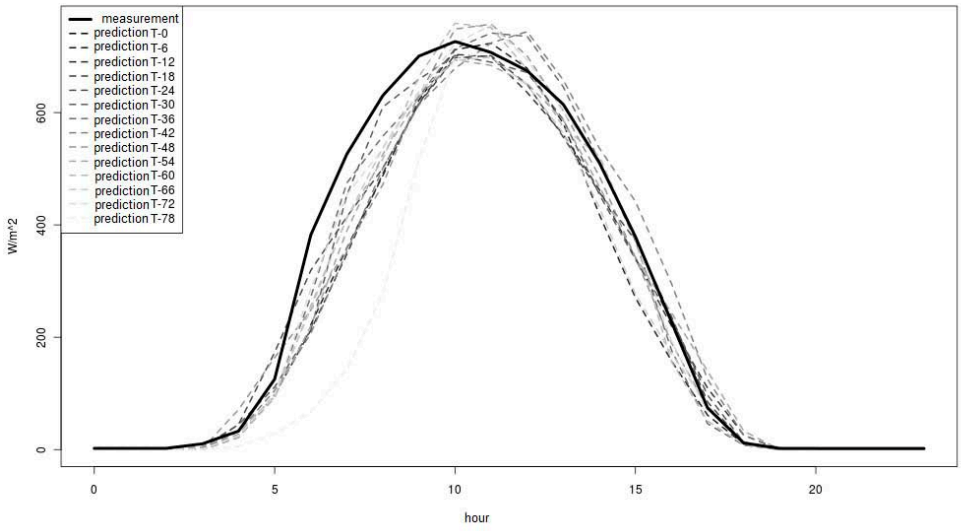


Fig. 13 *Time-lagged ensemble forecast of solar power in astronomic observatory of the Astronomic Institute of the Academy of Sciences of the Czech Republic in Ondřejov for day May 15, 2009.*

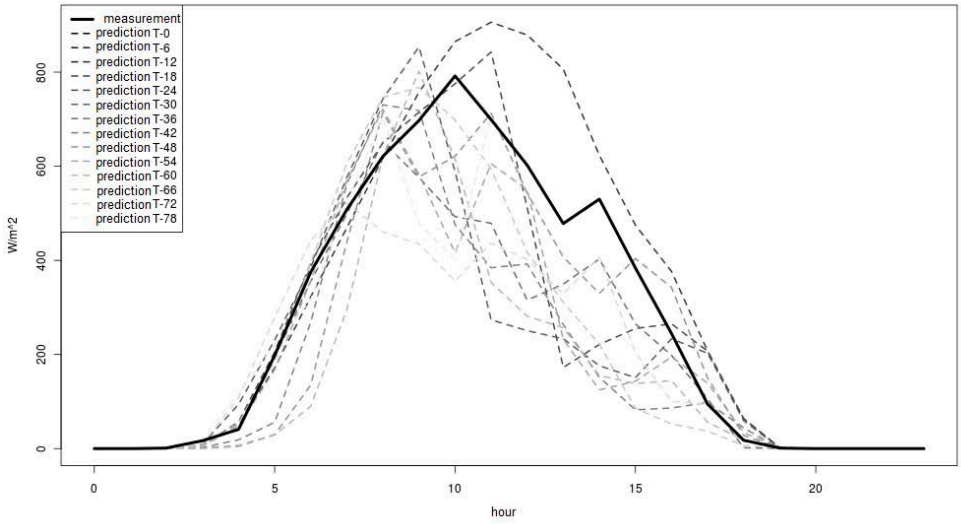


Fig. 14 *Time-lagged ensemble forecast of solar power in astronomic observatory of the Astronomic Institute of the Academy of Sciences of the Czech Republic in Ondřejov with increased cloudiness in the afternoon period, predicted by numerical weather prediction model MM5 for day June 10, 2008.*

reflected while increased cloudiness in the afternoon in Fig. 14. So the time-lagged ensemble predictions can be a substitute for compute-intensive classical ensemble predictions.

10. Final Remark

Interest in the use of forecasting methods in many areas of human activity is still very large. Forecasting techniques are constantly improving in accordance with the theoretical and practical knowledge and the development of computer technology, which enables the implementation of previously difficulty feasible experiments and lengthening the forecast horizon with practical usability. Forecasts provided by sophisticated mathematical models have become part of everyday life. The big challenge for prediction models developers is the more precise quantification of uncertainties in forecasts, and in particular its more understandable interpretation for the common user.

Curriculum Vitae

Emil Pelikán is a Professor at the Czech Technical University in Prague. He received his PhD degree in Computer Science from the Institute of Computer Science, Academy of Sciences of the Czech Republic in 1985. He was appointed to the position of Associate professor in the Faculty of Transportation Sciences at the Czech Technical University in Prague in 2002. His research interests includes time series modeling and prediction, statistical data analysis and data mining method. He was principal investigator of many grants and projects concerning of development of statistical and artificial intelligence methods in air pollution, energy sectors, meteorology, transportation area, banking and medicine. He published over 150 publications including journal and conference papers and books.

References

[1] BISHOP C.M. *Neural networks for Pattern Recognition*. Oxford: Oxford University Press, 1995.

[2] BOCCARA N. *Modeling Complex Systems*. New York: Springer, 2010, doi: 10.1007/978-1-4419-6562-2.

[3] DURBIN J., KOOPMAN S.J. *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press, 2004, doi: 10.1093/acprof:oso/9780199641178.001.0001.

[4] EBEN K., JURUŠ P., RESLER J., BELDA M., PELIKÁN E., KRUEGER B.C., KEDER J. *An ensemble Kalman filter for short-term forecasting of tropospheric ozone concentrations*. Q.J.R. Meteorol. Soc. Part C. 2005, 131(613), pp. 3313–3322, doi: 10.1256/qj.05.110.

[5] EBEN K., JURUŠ P., RESLER J., BELDA M., KRÜGER B.C., KEDER J. *MEDARD – An Environmental Modelling Project for the Territory of the Czech Republic*. ERCIM News, 61, 2005.

[6] FISCHER J. *Macroeconomic Forecasts: Betraying Us When Most Needed*. International Symposium on Forecasting, Prague, 2011. Keynote speech.

[7] GRABEC I., SACHSE W. *Synergetics of Measurement, Prediction and Control*. Berlin: Springer Verlag, 1997, doi: 10.1007/978-3-642-60336-5.

[8] JURUŠ P. *Modelování atmosféry v regionálním měřítku s aplikacemi v dopravě*. Prague, 2009. PhD thesis, Czech Technical University, Faculty of Transportation Sciences.

[9] KALNAY E. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge: Cambridge University Press, 2004, doi: 10.1017/cbo9780511802270.

[10] LORENZ E.N. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*. 1963, 20, pp. 130-141, doi: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

[11] MAKRIDAKIS S., WHEELWRIGHT S.C., HYNDMAN R.J. *Forecasting, Methods and Applications*. 3rd ed. New York: John Wiley & Sons, 1998.

[12] *Medard On-Line*. Institute of Computer Science of Academy of Sciences of the Czech Republic, 2003 [viewed 2014-12-18]. Available from `http://www.medard-online.cz`.

[13] NICOLIS G., NICOLIS C. *Foundations of Complex Systems*. Singapore: World Scientific, 2007, doi: 10.1142/9789812775658.

[14] PELÁNEK R. *Modelování a simulace komplexních systémů*. Brno: Masaryk University Press, 2011.

[15] PELIKÁN E., DE GROOT C., WÜRTZ D. Power Consumption in West-Bohemia: Improved Forecasts with Decorrelating Connectionist Networks. *Neural Network World*, 1992, 6(2), pp. 701–712.

[16] PELIKÁN E. *Processes Forecasting in Complex Systems*. Prague, 2012. Lecture in front of the Scientific Council of the Czech Technical University in Prague, in Czech.

[17] PEÑA D., TIAO G.C., TSAY R.S. *A course in Time series Analysis*. New York: John Wiley & Sons, 2001.

[18] RAIDL A. O (ne)předpověditelnosti počasí. *Pokroky matematiky, fyziky, a astronomie*. 1997, 42(6), pp. 302–312.

[19] RESLER J. *Inverzní modelování emisí z dopravy*. Prague, 2009. PhD thesis, Czech Technical University, Faculty of Transportation Sciences.

[20] SCHOLKOPF B., SMOLA A.J. *Learning with Kernels, Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press, 2002.

[21] TAKENS F. Detecting strange attractors in turbulence. In: D.A. Rand, L.S. Young, eds. *Dynamical Systems and Turbulence*, Lecture Notes in Mathematics 898. Berlin: Springer-Verlag, 1981, pp. 366–381.