SELF-TUNING RECURSIVE MODELLING AND ESTIMATION OF WEATHER MEASUREMENTS

Vesa Hasu, Control Engineering Laboratory, Helsinki University of Technology, Finland, vesa.hasu@hut.fi
Jani Poutiainen, Finnish Meteorological Institute, Finland, jani.poutiainen@fmi.fi
Heikki Koivo, Control Engineering Laboratory, Helsinki University of Technology, Finland, heikki.koivo@hut.fi

Abstract. This paper describes a methodology for self-tuning modelling and estimation of meteorological measurements. The idea is to utilize recursive least squares (RLS) method in modelling of the weather measurement and use celebrated Kalman techniques for smoothing the possible erroneous data and estimation for missing measurement points. In addition to the normal single measurement problem, we will attack also the problem with multiple measurement stations in meteorological mesoscale network with average station distances in the order of 10 km.

1. Introduction

The state of the weather has been, is and will be one of the most popular discussion topics in the history of mankind. In addition to idle conversations, knowing the weather and reacting accordingly can have impact from everyday convenience to enormous economical savings.

For everyday safety and life, people, traffic operations, road and real estate maintenance and other businesses need information on the short-term local weather. Nowcasting refers to very short term weather forecasting. It is usually considered in 0-6 hours timeframe and in few kilometers spatially. Some of the typical phenomena include thunderstorms, fog, temperature inversions and sea breeze. The recent development of advanced and inexpensive measurement equipment has made possible to measure meteorological phenomena densely in large weather station networks with short measurement interval, thus offering new possibilities in weather analysis and nowcasting. Observations are necessary prerequisites for weather forecast models.

New types of measurement networks, and temporally and spatially increased amount of data bring up also challenges. The incoming in-situ surface observation data flow can easily be in different order of magnitude than before. Hence the first problems are faced in data preprocessing, the old quality control techniques may be insufficient or may even not be valid anymore. The final data quality control of all differing measurements has traditionally been left to meteorologists [4]. The requirement for human process supervision is essential also in future, but the level of quality controlled points should be kept at minimum with new algorithms and applications. In automated quality control, traditionally the simplest and widely used single station methods have included various limit checks and parameter cross-comparisons.

The new public real time data networks are likely to suffer from missing measurement values much more than old networks. On the other side, these networks allow higher utilization of advanced control engineering techniques. The request for meteorological metadata is steadily increasing. The measurement values alone are incomplete. Indicators of the measurement quality and uncertainty must accompany the actual measurement values. This type of metadata can be readily deduced from above mentioned quality assurance methods.

Meteorological observation networks are likely to develop towards autonomous dynamic data processes. For instance, there is need for adaptive data collection strategies depending on the weather patterns but also a requirement to incorporate automatic fault detection methods deeper into maintenance processes. Since measurement stations are everywhere, wireless links are required and they can sometimes drop measurements. It is also not uncommon instruments to be broken by a lightning or a bird. Some devices are sensitive to icing or radio frequency disturbances. Sometimes appropriate care of the instruments may be neglected, an example being misplaced shading devices in solar radiation measurements. Error sources are numerous and all the above incidents result errors in meteorological information. Sudden fault maintenance actions are costly, which is further emphasized in hard to reach and harsh environments like communication masts or lighthouses at the sea. Timely detection and analysis of severity of problems and corresponding selection of the most efficient maintenance decisions make a significant difference in daily operations and economically.

2. Measurement Network: Helsinki Testbed

The measurements utilized in this paper are from the Helsinki Testbed -mesoscale measurement network [1] which is led by Finnish Meteorological Institute and Vaisala measurements company. Outcomes from testbeds are more effective observing systems, better use of data in forecasts, improved services, products, and economic/public safety benefits. Testbeds accelerate the translation of research and development findings into better operations, services, and decision-making. The results of this study directly serve the idea behind a testbed in refining quality assurance techniques.

In 2005, the existing Finnish weather observation network was supplemented with nearly 60 stations equipped with Vaisala WXT510 weather transmitters. Of these stations, 42 consist of cellphone base station masts, converted to meteorological towers by installing weather transmitters on them at least at two levels per mast. A map of Testbed area is shown in Fig. 1. The measurements considered in this study are made by Vaisala WXT510 weather stations in the coastal part of southern Finland.

WXT510 stations measure six common weather parameters: wind speed and direction, liquid precipitation, barometric pressure, temperature and relative humidity. Pictures of WXT510 and its dimensions are presented in Fig. 2. In Helsinki Testbed, the measurement stations are located about 10 km apart from each other on the average. The measurements are recorded and collected in every 5 minutes. Measurements are transmitted in real time to the central database typically through GPRS connection. The system makes use of XML language and

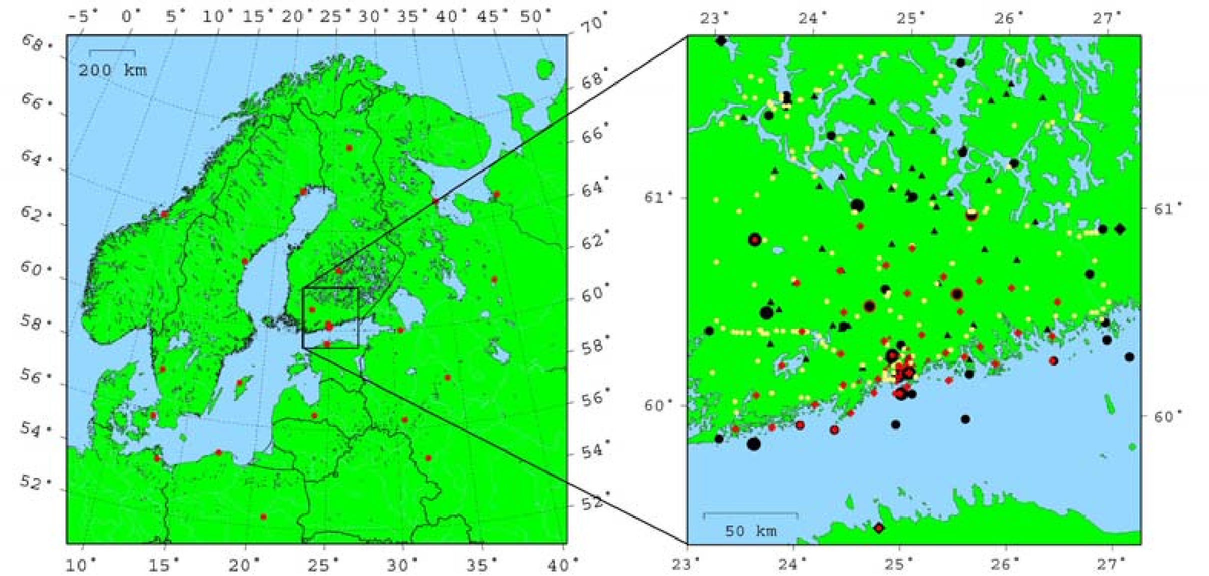

Fig. 1: Helsinki testbed observing sites in the inner domain (right) and outer domain (left). The data considered in this paper are collected in the points marked by small red diamonds.

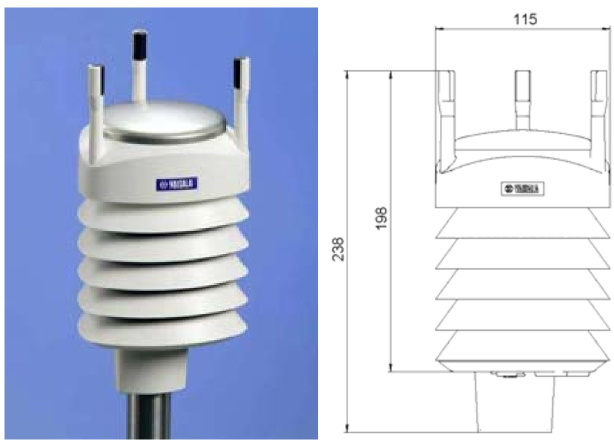

Fig. 2: Vaisala WXT510 weather station (left) and its dimensions in mm (right)

multi-tier service oriented architecture. Software layers exist on top of the database for data distribution to the Internet and mobile platforms.

3. The Modelling Motivation

The weather station network has three basic needs for the modelling and estimation. Firstly, the possible noisy measurements can be easily filtered. Such noise can occur due to sensor flaws, a poor installation or weather conditions. Especially weather conditions make the filtering interesting. Distinguishing actual weather phenomena from noise is a difficult task for filtering and estimation.

The second need for modelling is the estimation of missing measurements. The absence of measurements can occur *e.g.* due to wireless link mishap, maintenance break or sensor failure. The estimation of missing measurements can be done according to the same measurement history or according to the measurements of the neighbor stations.

Thirdly, data quality indicators such as estimation error are needed. Any additional data describing the primary parameters or observation conditions can be called meteorological metadata, which in turn may be used to create combined quality indicators, often referred to as data flags.

To summarize, practical and important end-user application is the detection of non-meteorological signals and production of data quality indicators for all data.

The scope of the current research is concentrated on the measurements and their quality, not so on the predicting values of meteorological measures for future time steps. Our approach of time series modelling is not accurate at all through the whole nowcasting timeframe of several hours.

4. Mathematical Background

The weather measurements are modelled as moving average time series model. The model is updated in each iteration by a recursive least squares method (RLS) [2]. This model is formatted into state-space representation and used in Kalman-techniques for estimation and prediction purposes. Kalman-techniques are widely covered in the literature, *e.g.* [3].

The Kalman-predictor can compensate the missing measurement values and the Kalman-estimator can smooth out possible excessive noise. The covariance matrices for system and measurement noise required for Kalman filter are estimated recursively. This guarantees the applicability of the model for different measurements in different meteorological conditions, *e.g.* measurements in inland and coastal stations can have very different statistical properties.

The missing measurements can decrease the system performance significantly, if the measurement is missing for a long period and the neighbor stations are far away. A long period with no measurements and the model structure leads to an autoregressive sequence, which is not accurate after a few hours of measurements. Hence additional parallel model structures are applied, in which the missing measurement is removed in RLS model.

4.1. Recursive Least Squares Modelling

The recursive least squares (RLS) approach is based on moving average prediction model of vector $\mathbf{x}$ of dimensions n x 1

$$\hat{\mathbf{x}}(k+1 \mid k) = \left(\mathbf{A}_0(k) + \mathbf{A}_1(k)q^{-1} + ... + \mathbf{A}_m(k)q^{-m}\right)\mathbf{x}(k) \tag{1}$$

where q^{-1} is the delay operator for difference systems (for more about delay operator, see *e.g.* [5]), and **A**-matrices are of dimension n x n. The notation $(k + 1 \mid k)$ stands for the estimate of time step $k + 1$, while data until step k is known. If we write

$$\mathbf{X}(k) = \left[\mathbf{x}^T(k) \quad ... \quad \mathbf{x}^T(k-m)\right]^T, \tag{2}$$

and

$$\mathbf{A}(k) = \left[\mathbf{A}_0(k) \quad \dots \quad \mathbf{A}_m(k)\right], \tag{3}$$

then the model (1) can be expressed as

$$\hat{\mathbf{x}}(k+1 \mid k) = \mathbf{A}(k)\mathbf{X}(k). \tag{4}$$

In this case, the parameter matrix **A** can be estimated by RLS model [2]:

$$\boldsymbol{\gamma}(k) = \frac{\mathbf{C}(k)\mathbf{X}(k)}{\mathbf{X}^T(k)\mathbf{C}(k)\mathbf{X}(k) + \beta}, \tag{5}$$

$$\mathbf{A}(k+1) = \mathbf{A}(k) + \boldsymbol{\gamma}(k)\left(\mathbf{x}(k+1) - \mathbf{A}(k)\mathbf{X}(k)\right), \tag{6}$$

$$\mathbf{C}(k+1) = \left(\mathbf{I} - \boldsymbol{\gamma}(k)\mathbf{X}(k)\right)\mathbf{C}(k)/\beta, \tag{7}$$

where $\gamma(k)$ is the time-variable update gain (sometimes referred to as the Kalman gain), $\mathbf{C}(k)$ is the inverse of the covariance matrix of the variables in $\mathbf{X}(k)$, β is the forgetting factor, and **I** is a unit matrix with suitable dimensions. The forgetting factor β is chosen from the range (0, 1]. Its name is due to the fact that the weight of the older data points decay with the factor β. Usually it is chosen to be close to 1.

4.2. Kalman Filtering

The Kalman filtering is based on the state-space representation of the system. In our case, the model (2) must be converted accordingly. Hence the following state-space representation is obtained

$$\begin{cases} \mathbf{X}(k+1) = \boldsymbol{\Phi}(k)\mathbf{X}(k) + \mathbf{w}(k) \\ \quad \mathbf{x}(k) = \mathbf{H}\mathbf{X}(k) + \mathbf{v}(k) \end{cases} \tag{8}$$

where **w** and **v** are independent unknown process and measurement noises with $\mathbf{w}(k) \sim N(0,\mathbf{Q}(k))$, $\mathbf{v}(k) \sim N(0,\mathbf{R}(k))$, $\boldsymbol{\Phi}(k)$ is the state transition matrix between time steps k and $k + 1$, and **H** is the observation matrix. In our case, matrices $\boldsymbol{\Phi}$ and **H** are defined by

$$\boldsymbol{\Phi}(k) = \begin{bmatrix} \mathbf{A}(k) \\ \mathbf{0}_{nmxn} \quad \mathbf{I}_{nm} \end{bmatrix} \text{ and } \mathbf{H} = \left[\mathbf{I}_n \; \mathbf{0}_{nxmn}\right], \tag{9}$$

where $\mathbf{0}_{nxm}$ is a $n \ x \ m$ zero matrix, and $\mathbf{I}_n$ is a $n \ x \ n$ unit matrix.

Now, the initial conditions and assumptions for Kalman filter are

$$E\{\mathbf{X}(0)\} = \hat{\mathbf{X}}(0), \; E\left\{\left(\mathbf{X}(0) - \hat{\mathbf{X}}(0)\right)\left(\mathbf{X}(0) - \hat{\mathbf{X}}(0)\right)^T\right\} = \mathbf{P}(0) \text{ and } E\left\{\mathbf{w}(k)\mathbf{v}(j)^T\right\} = \mathbf{0}, \forall \, k, j. \tag{10}$$

By definition, $\mathbf{P}(k)$ is the covariance matrix of the estimation error in the time step k. The prediction equations for the Kalman filter are

$$\hat{\mathbf{X}}(k+1 \mid k) = \boldsymbol{\Phi}(k)\hat{\mathbf{X}}(k \mid k), \tag{11}$$

$$\mathbf{P}(k+1 \mid k) = \boldsymbol{\Phi}(k)\mathbf{P}(k \mid k)\boldsymbol{\Phi}(k)^T + \mathbf{Q}(k), \tag{12}$$

and the Kalman filter update equations are

$$\mathbf{K}(k+1) = \mathbf{P}(k+1 \mid k)\mathbf{H}^T\left(\mathbf{H}\mathbf{P}(k+1 \mid k)\mathbf{H}^T + \mathbf{R}(k)\right)^{-1}, \tag{13}$$

$$\mathbf{P}(k+1 \mid k+1) = \left(\mathbf{I} - \mathbf{K}(k+1)\mathbf{H}\right)\mathbf{P}(k+1 \mid k), \tag{14}$$

$$\hat{\mathbf{X}}(k+1 \mid k+1) = \hat{\mathbf{X}}(k+1 \mid k) + \mathbf{K}(k+1)\left(\mathbf{x}(k+1) - \mathbf{H}\hat{\mathbf{X}}(k+1 \mid k)\right). \tag{15}$$

5. Practical Implementation Details

The first problem in the application of the above Kalman filter in practice is the estimation of noise covariance matrices. Since there are no better estimates available, the process noise covariance matrix **Q** and the observation noise matrix **R** are estimated from the increments in sequential state estimates and observations, respectively. Mathematically that is

$$\mathbf{Q}(k+1) = \lambda_q \mathbf{Q}(k) + (1-\lambda_q)\left(\left(\hat{\mathbf{X}}(k+1|k+1) - \hat{\mathbf{X}}(k|k)\right)\left(\hat{\mathbf{X}}(k+1|k+1) - \hat{\mathbf{X}}(k|k)\right)^T\right), \tag{16}$$

$$\mathbf{R}(k+1) = \lambda_r \mathbf{R}(k) + (1-\lambda_r)\left(\left(\mathbf{x}(k+1) - \mathbf{x}(k)\right)\left(\mathbf{x}(k+1) - \mathbf{x}(k)\right)^T\right), \tag{17}$$

where λ_q and λ_r are forgetting factors.

In practice, the largest problem faced by the above RLS-KF combination is the missing observations. A solution is to replace the possible missing observations with estimated observations. However, if the measurement is missing for a long period, the estimation becomes eventually autoregressive and the levels of all estimated observations in $\hat{\mathbf{X}}$ become biased. In our weather station case, the measurements may be missing for several days (or for hundreds of iterations). Hence an alternative solution is required, and we propose using parallel partial RLS models.

The use of parallel partial RLS models is based on not using the full independent (input) data in modelling of full or partial dependent (output) data. Here independent data refers to $\mathbf{X}(k)$ and dependent data refers to $\mathbf{x}(k+1)$. In our RLS model update approach, if one or more of the observation in m previous steps is missing, models with that measurement on the independent side of the model (4) are not updated. If observation is missing in the dependent side, the corresponding row of **A** is not updated. In the Kalman estimation, a measurement is ignored in the dependent side only if all previous $m + 1$ observations are missing. Fig. 3 is demonstrating which kind of models can be updated and what kind of model can be used in estimation. The example includes four measurements with some missing observations.

A problem of using parallel RLS models is the exponentially growing number of models in respect to dimension n (number of parallel RLS models to update is $2^n - 1$, to be exact). However, in our weather station case, the model is intended to be applied in the physical neighborhood of a single station. Therefore the dimension n does not usually grow larger than 6 or 7.

Nevertheless, if the estimation system is applied in large scale, it is advisable not to update all models in each possible round. For example in our simulation cases, all RLS models with at most one missing input measurement are updated every possible step. The models with more missing input measurements are divided into three subgroups. Thus in our example of neighborhood of five stations the average number of simultaneously updated parallel models is about 14.

6. Examples

All the following examples are taken from a five weather station set near Helsinki area in the southern coast of Finland. A map including approximate positions of weather stations is drawn in Fig. 4. RLS model time window is three previous measurements and RLS forgetting factor β is 0.9. The error covariance estimation forgetting factors λ_q and λ_r are both set to 0.95.

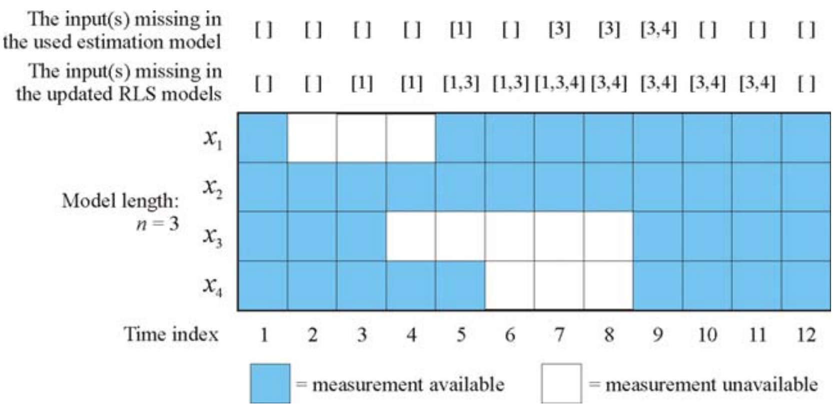

Fig.3: A sketch of applied inputs, while some measurements in **x** are missing.

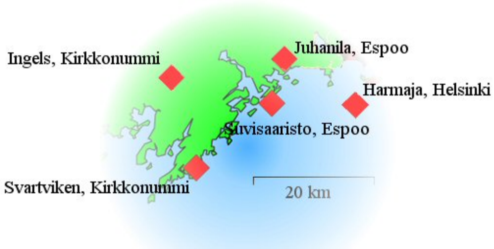

Fig. 4: Measurement stations in the southern Finland.

6.1 Filtering

The first example of filtering effects is given in Figs. 5, which includes unfiltered and filtered temperature values of the five stations mentioned in above in September 1st 2005. The measurements are not that noisy and therefore the filtering is not strong. The breaks in the filtering difference graphs are due to the missing observations. Fig. 6 includes an example, when the same case is filtered with some additive noise in the measurement of Suvisaaristo's temperature. In compared to the original measurement, the mean square errors of the signal with noise and filtered signal are 3.78 and 2.35.

The examples give an impression that the filtering effect is not large with the above parameter choices. However, if filter was smoothing the signal more, there is a large danger of filtering out also genuine meteorological phenomena – which certainly is not desired. As an example of this difficulty, Fig. 7 has a case of a barometric pressure, in which the presented filtering is applied. In Fig. 7a, Svartviken measurement has a measurement error in iteration 7142, which is not filtered completely out. However, Fig. 7a has also a rapid barometric pressure change after iteration 7250, which would have been smoothed out in heavier filtering, see Fig. 7b (the change can be identified as a meteorological phenomena, since the other stations also has the same character). The heavier filtering is implemented through setting RLS forgetting factor β to 0.99.

6.2 Estimation of Missing Observations

The accuracy of the estimation of missing observations depends on a few factors. The first factor obviously is the level of noise in the measurements. The second factor is the level of correlation between the measurements on the input side of the model. Additionally, the estimation is dependent on the number of simultaneously missing observations, the length of the period of values missing, and the effective area of the meteorological phenomena. These factors cause large differences in the estimation accuracy of missing observations. In the following, barometric pressure and temperature observations of Suvisaaristo station are estimated based on the neighboring stations (see Fig. 4).

An example of estimated barometric pressure is drawn in Fig. 8, where is plotted air pressure of Suvisaaristo station. The first 3000 observations are measured and filtered, but the following 4000 points are estimated based on the four stations in the neighborhood. The pressure difference drawn in Fig. 8b indicates that the estimation between approximately iterations 3000-7000 does not reduce the accuracy that much. If the difference is estimated in mean square error, the estimation increases the difference in four-fold.

The estimation accuracy depends on the nature of meteorological quantities. Examples of estimated temperatures at the Suvisaaristo station are plotted in Fig. 9. Figures show that the estimated temperature is biased and does not present the current weather change (Fig. 9a), the estimation can have a large error, but present the nature of phenomenon (Fig. 9b), or the estimation can be on fairly accurate even for a long time (Figs. 9c and 9d, in which the estimated period is 14 and 15 hours, respectively).

In optimization of the model use, it might be beneficial to adjust model conditions for each meteorological parameter separately, and according to time, season, location, topography, and prevailing weather. One simpler option to reduce the overall temperature estimation error would be to introduce time-of-day –dependency into RLS models. However, even this approach would suffer from the large memory requirement – the number of models would have to be multiplied with 288.

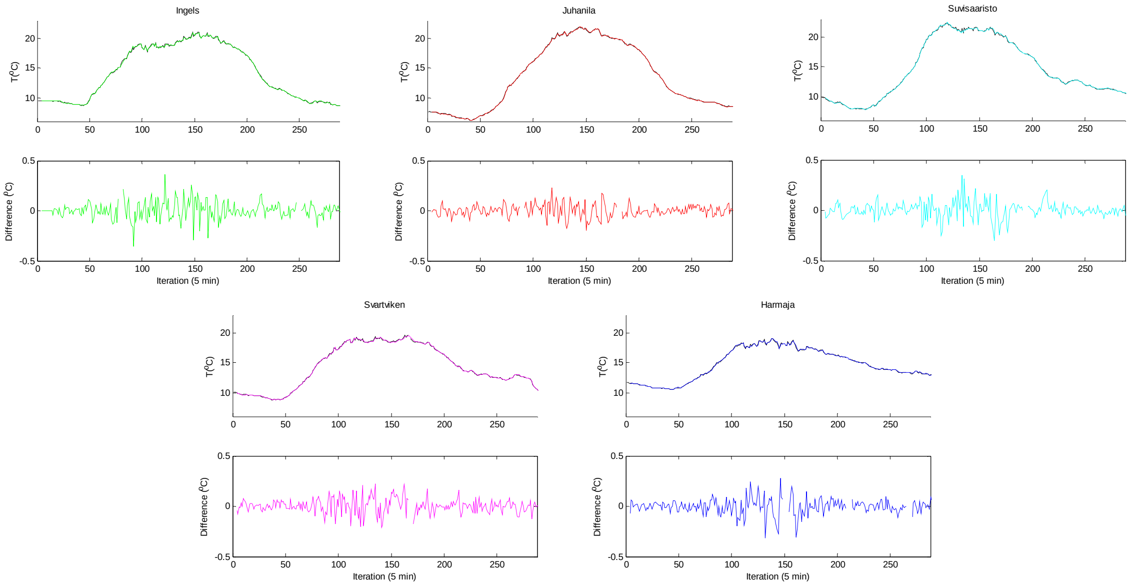

Fig. 5: An example of temperature filtering. The original observations are drawn in black color.

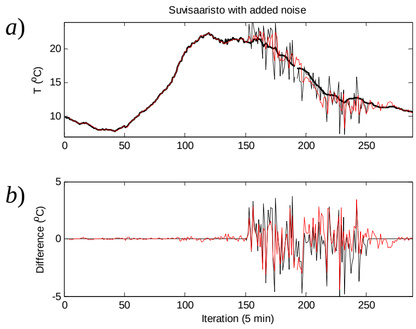

Fig. 6: An example filtering of temperature with noise, *a*) thick black line is the correct measurement, thin black line is the measurement with noise and the red one is the filtered signal, *b*) the differences to the original measurement.

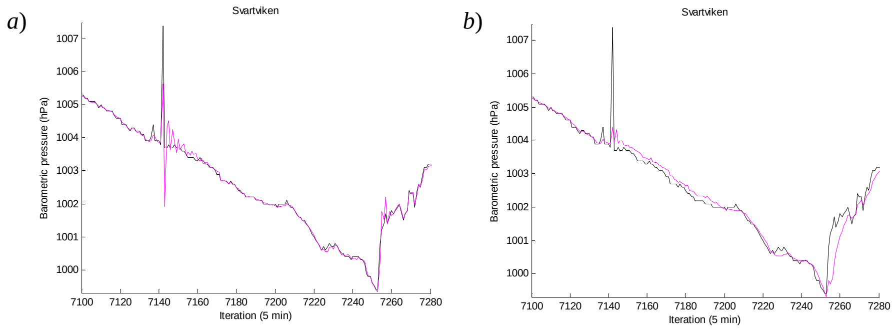

Fig. 7: An example of the difficulty of weather measurement filtering, barometric pressure of Svartviken station is plotted with light and heavy smoothing in *a*) and *b*), respectively.

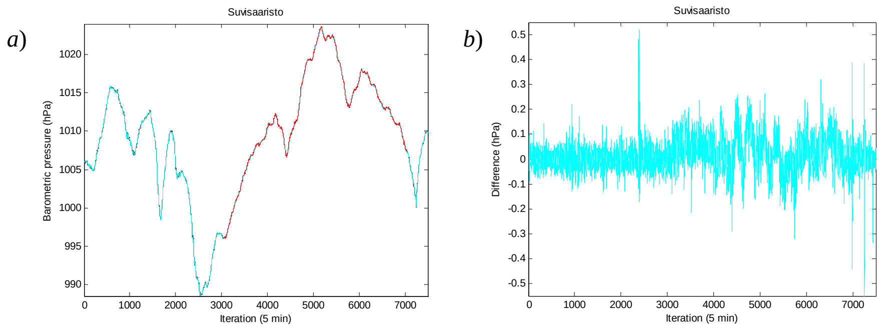

Fig. 8: An example of *a*) missing value estimation in the case of barometric pressure and *b*) the difference between the estimated and measured value.

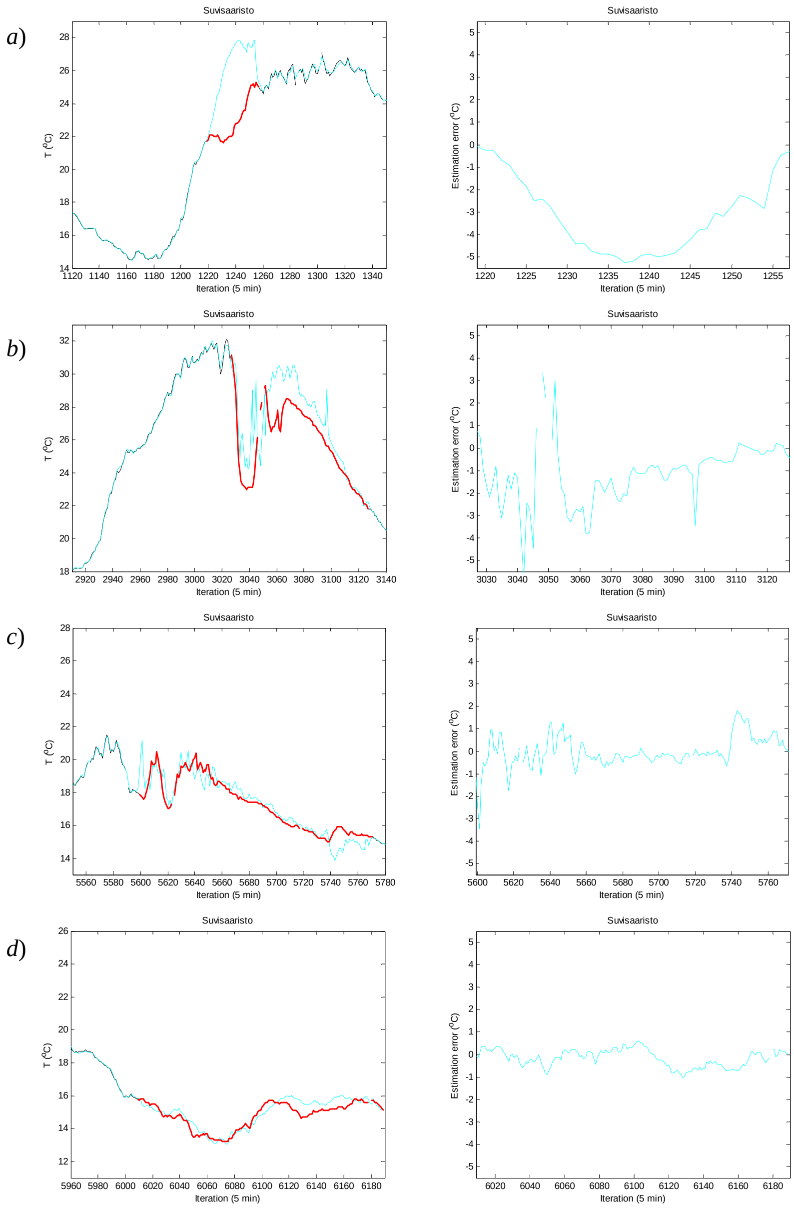

Fig. 9: Examples of estimation of missing observations. Estimates and missing observations are drawn in blue are red, respectively. On right, *x*-axes are extended to cover only the periods of missing data.

7. Conclusions

All in all, recursive modelling with Kalman filter seems to be a promising technique for modelling, filtering, and estimation of weather measurements. In this approach, the filtering based on the RLS models can be tuned with the forgetting factor of the RLS update. In the filtering must be kept mind that all variations may not be noise and therefore they must not be filtered out completely. The success of estimation of missing observations varies a lot. Estimation accuracy depends on meteorological and measurement history factors. On the considered variables, barometric pressure seems to be easily estimated based on the neighboring stations, whereas temperature estimation proves to be tricky at times. These findings are in good agreement with meteorological nature of the corresponding parameters.

In addition to the presented filtering and estimation, the underlying purpose of presented RLS-KF –combination is to identify faulty measurements coming from weather stations. The filtered signal provides a residual, which can be utilized in fault detection. In the end, the practical aim is to formulate new quality assurance methods to be incorporated into operational meteorological information production processes.

References

[1] W. Dabberdt, J. Koistinen, J. Poutiainen, E. Saltikoff, H. Turtiainen, “The Helsinki Mesoscale Testbed - An Invitation to Use a New 3-D Observation Network”, *Bulletin of the American Meteorological Society*, 86(7), July 2005, pp. 906-907.

[2] J. Holst, *Adaptive Prediction and Recursive Estimation*, Department of Automatic Control, Lund Institute of Technology, Sweden, 1977.

[3] G.C. Goodwin, K.S. Sin, *Adaptive Filtering Prediction and Control*, Prentice-Hall, NJ, USA, 1984.

[4] F. Vejen (ed.), C. Jacobsson, U. Fredriksson, M. Moe, L. Andersen, E. Hellsten, P. Rissanen, Þ. Pálsdóttir, Þ. Arason, *Quality Control of Meteorological Observations*, Nordklim report, 2002, http://www.smhi.se/hfa_coord/nordklim/task1/quality_control.pdf.

[5] K.J. Åström, B. Wittenmark, *Computer Controlled Systems*, Theory and Design, 3rd ed., Prentice Hall, 1997.