Assimilation of lake water surface temperature observations using an extended Kalman filter

By EKATERINA KOURZENEVA[1,2]*, [1]*Finnish Meteorological Institute, P.O. Box 503, FI-00101, Helsinki, Finland*; [2]*Russian State Hydrometeorological University, Malookhtinsky pr., 98, RU-195196, Saint-Petersburg, Russia*

ABSTRACT

A new extended Kalman filter (EKF)-based algorithm to assimilate lake water surface temperature (LWST) observations into the lake model/parameterisation scheme Freshwater Lake (FLake) has been developed. The data assimilation algorithm has been implemented into the stand-alone offline version of FLake. The mixed and non-mixed regimes in lakes are treated separately by the EKF algorithm. The timing of the ice period is indicated implicitly: no ice if water surface temperature is measured. Numerical experiments are performed using operational in-situ observations for 27 lakes and merged observations (in-situ plus satellite) for 4 lakes in Finland. Experiments are analysed, potential problems are discussed, and the role of early spring observations is studied. In general, results of experiments are promising: (1) the impact of observations (calculated as the normalised reduction of the LWST root mean square error comparing to the free model run) is more than 90% and (2) in cross-validation (when observations are partly assimilated, partly used for validation) the normalised reduction of the LWST error standard deviation is more than 65%. The new data assimilation algorithm will allow prognostic variables in the lake parameterisation scheme to be initialised in operational numerical weather prediction models and the effects of model errors to be corrected by using LWST observations.

Keywords: land surface data assimilation, parameterisation of lakes, extended Kalman filter

1. Introduction

In recent years, it was demonstrated that lakes affect local weather conditions and regional climate (Krinner, 2003; Eerola et al., 2010; Samuelsson et al., 2010). With the increasing horizontal resolution of numerical weather prediction (NWP) and climate models to the kilometre scale, more and more attention is being paid to the interaction between the atmosphere and lakes. Over lakes, turbulent and radiative fluxes are dependent on the water surface temperature and the presence of ice (Mironov, 2008; Rontu et al., 2012). Their influence is essential for countries with a high percentage of lakes, such as Finland, Sweden, Canada, Norway and Russia.

To account for the effects of lakes in atmospheric models, physically sound and computationally cheap lake models (parameterisation schemes) are used. The most commonly used lake parameterisation scheme is the lake model Freshwater Lake (FLake) (Mironov, 2008), with a two-layer parametric representation of the water temperature profile. External parameters for the lake scheme are provided by the Global Lake Database (Kourzeneva et al., 2012a; Choulga et al., 2014). For NWP initialisation at the very first simulation, lake climatology is developed (Batrak, 2012; Kourzeneva et al., 2012b). FLake is widely used in NWP and climate modelling for research (Salgado and Le Moigne, 2010; Samuelsson et al., 2010; Balsamo et al., 2012; Martynov et al., 2012). It is also included in operational NWP model runs in some National Weather Service centres (Mironov et al., 2012; Rontu et al., 2012). However, it has not become a standard operational practice in NWP. There are several reasons for this, including too simplified data assimilation (DA) methods (borrowed from ocean surface state representation). Through initialisation of the prognostic lake model variables, DA techniques should provide the analysis with corrected errors which come from an unknown initial state. Currently used DA techniques do not fulfil this task, although model errors for the lake surface state may be quite large in NWP (Batrak, 2012; Kourzeneva et al., 2012b).

*Correspondence.
email: Ekaterina.Kurzeneva@fmi.fi
Responsible Editor: Nils Gustafsson, SMHI, Norrköping, Sweden.

1

In-situ or remote-sensing observations of different water quantities are a significant source of information about the state of lakes. Observations of the lake water surface temperature (LWST) and ice cover are mostly important for NWP and climate modelling since these values control energy and moisture exchanges between the atmosphere and lake surface [see, e.g., Rontu et al. (2012)]. Space-borne observations contain information about LWST and ice cover. They cover large territories and have spatial resolutions from kilometres to tens of metres and temporal resolutions of several hours. However, they suffer from coverage gaps and different errors, for example, from cloud contamination or skin effects (Kheyrollah Pour et al., 2014a). In-situ data may contain measurements of different physical characteristics of water and ice, for example, the ice depth and the temperature profiles in water, ice and snow. They contain fewer errors compared with satellite observations. But they are very sparse and produced mainly for research purposes. On an operational basis, the LWST of only about 30 lakes in Finland is measured in-situ by the Finnish Environmental Institute (SYKE) [see Kheyrollah Pour et al. (2014b), Rontu et al. (2012) and Eerola et al. (2010) for details].

To combine optimally information from a model with observations, DA methods may be used. Experiments performed in Kheyrollah Pour et al. (2014b), Rontu et al. (2012) and Eerola et al. (2010), where different LWST and ice cover observations were treated by the NWP model in quite a simplified way, showed that assimilation of LWST observations may improve the description of the lake surface state and the quality of a weather forecast. For example, in a winter case study over Lake Ladoga, a better description of ice cover due to the assimilation of remote-sensing data reduced the screen level temperature forecast errors by up to 5°C for several selected stations in Finland (Kalle Eerola, personal communications 2014). DA systems are widely used in meteorology (Kalnay, 2003), in land surface modelling (Houser et al., 2010) and in oceanography (Haines, 2010). In hydrology or limnology, there are some studies devoted to DA for inland water bodies (Zhang et al., 2007; Stroud et al., 2009). However, the focus of their research is confined to specific lakes. They concentrate mainly on circulation characteristics in one large lake, using 3D lake models and DA techniques used for oceanography. Unfortunately, 3D lake models cannot be used to parameterise all lakes in an atmospheric model domain which differ in size, form and depth. Hence, oceanographic techniques are not applicable for lakes in NWP. Although lakes belong to the hydrosphere and lake modelling is close to oceanography, the atmosphere 'sees' the lake surface as part of a very heterogeneous land surface. Since simplified 1D or even bulk lake models are used for parameterisation purposes in NWP, methods used to assimilate LWST observations

should be closer to those applied in the 1D aspect of land surface DA (Mahfouf, 1991; Rhodin et al., 1999; Hess, 2001; Mahfouf et al., 2009; de Rosnay et al., 2013).

Thus, assimilation of lake observations includes two aspects. The first aspect is to spread information in the horizontal direction, that is, to interpolate data from observational points or from the satellite image grid to the atmospheric model grid. This task is similar to the analysis of the screen level temperature, relative humidity and snow depth in land surface analysis, or to the 2D analysis of the sea surface temperature (SST) [see, e.g., Donlon et al. (2012), Homleid (2009), Brasnett (1999), Navascues (1997) and Gustafsson (1985)]. The main method used in this kind of analysis is optimal interpolation (OI) (Gandin, 1965). However, in contrast with sea water surface properties, lake water surface properties are very heterogeneous. Interpolation of LWST between two neighbouring lakes differing in size, depth and elevation is rather questionable. Perhaps the anisotropic surface conditions may be accounted for by using structure functions dependent not only on the separation in the horizontal (as for SST) but also on differences in depth and elevation. In practice this aspect also includes data quality control, processing of very dense satellite observations and questions of consistency between land-water masks. These problems are discussed in detail in Kheyrollah Pour et al. (2014b).

The second aspect is to propagate information about the lake surface state in the vertical direction, in the lake model space. In other words, to redistribute innovations (departures between observed and simulated values) of LWST in the lake water vertical profile (1D aspect). Mathematically, this task is analogous to the analysis of the soil moisture [see, e.g., Mahfouf et al. (2009) and Hess (2001)], and the same algorithms can be applied. In the case of soil moisture analysis, several methods of different complexity exist, starting from OI (Mahfouf, 1991) to variational methods and extended Kalman filtering (Rhodin et al., 1999; Hess, 2001; Mahfouf et al., 2009; de Rosnay et al., 2013). Nudging, which is also applicable, and OI are computationally cheap. Yet, their coefficients are dependent on an observational dataset to derive error statistics and on partition of the dataset according to physical situations. Also it is difficult to account for new observations. Variational methods and different versions of the extended Kalman filter (EKF) are more flexible for different physical situations and can accept new observations more readily. In Kalman filtering algorithms, the background error covariance matrix is prognostic, dependent on the physical situation, and no preliminary error statistics study is needed. However, these methods require greater computing resources. Some of these need additional simulations to compute Jacobians in finite differences. Balsamo et al. (2007) showed that the high computational cost may be reduced by running a

surface model offline, using saved forcings from the atmospheric model. In this case, variables from the lowest atmospheric model level are used to force additional runs (with the perturbed initial state) of the surface model. This approach is used, for example, to assimilate satellite measurements of the soil moisture (Draper et al., 2009). For lakes, this approach also could be used. The most common lake parameterisation scheme FLake contains a maximum of 12 prognostic variables, depending on the lake regime. Therefore the control vector size is small, and the EKF technique seems a natural one to apply.

The purpose of this study is to apply the EKF approach to assimilate LWST observations into the FLake lake model. For these first experiments, the offline version of the lake model was used, with forcing from the High Resolution Limited Area Model (HIRLAM) (Undén et al., 2002). SYKE in-situ observations were assimilated, for some lakes with added observations from the Moderate Resolution Imaging Spectrometer (MODIS) on board of Terra and Aqua satellites. In Section 2, the methodology of applying EKF in FLake (mathematical formulation) is described. Section 3 is devoted to lake observations. In Section 4, numerical experiments are presented and discussed, focusing on the role of observations in different seasons. Conclusions complete the paper.

2. Methodology

2.1. Lake model properties essential for DA

We assimilated observations into the lake model FLake (Mironov, 2008). This model is used to represent lake processes in several regional and global NWP models for both operational and research applications (Salgado and Le Moigne, 2010; Balsamo et al., 2012; Rontu et al., 2012 Mironov et al., 2012). It is a two-layer integral (bulk) model with the temperature profile in the upper mixed layer and in the underlying thermocline parameterised with the concept of self-similarity (assumed shape). It also contains snow-ice and bottom sediment modules using the same concept to describe their temperature profiles. The mixed layer depth is predicted through an equation of convective entrainment and a relaxation-type equation during the wind mixing. For the solar radiation transfer, the decay law is approximated exponentially. Using atmospheric forcings, turbulent fluxes in the atmospheric surface layer may also be calculated by the model. This possibility is useful for offline experiments.

To accurately pose the DA task for lakes (and for FLake in particular), it is important to consider different model regimes and to study the model behaviour for each of them. Firstly, we distinguish the ice period and the open water period. Secondly, in the open water period, we distinguish

the mixed and non-mixed (stratified) lake regimes. Thirdly, in the non-mixed period the convective and wind-mixing regimes may be discriminated.

- *Ice and open water periods*. In our study, the lake model runs freely during the ice period because there are no in-situ LWST observations available. However, we use the fact that if in-situ LWST observations exist, then there is definitely open water. In the future, for the assimilation of remote-sensing observations of LWST and of the lake ice surface temperature (LIST), it may be problematic to move smoothly from one surface type to another because of their different physics. For example, LIST evolves fast and has a very strong diurnal cycle, while LWST changes much slower, especially before freezing. Different model blocks are active in the different cases, and it is safer to assign different state vectors to them.
- *Mixed and stratified regimes*. As a rule, boreal lakes are dimictic (Lewis, 1983), that is, mixed in early spring and autumn and stratified in summer. This behaviour is well described by the lake model. For illustration, the typical annual cycle of the model mixed layer depth in a boreal lake is represented in Fig. 1. It is important to notice that jumps between regimes are quite rapid in the model. In these situations the model is strongly non-linear. Moreover, a model variable such as the shape factor (the integral of the polynomially approximated temperature profile in the thermocline, C_T) is even discontinuous when changing regimes (it does not exist in the mixed mode; see Section 2.2 for more details).
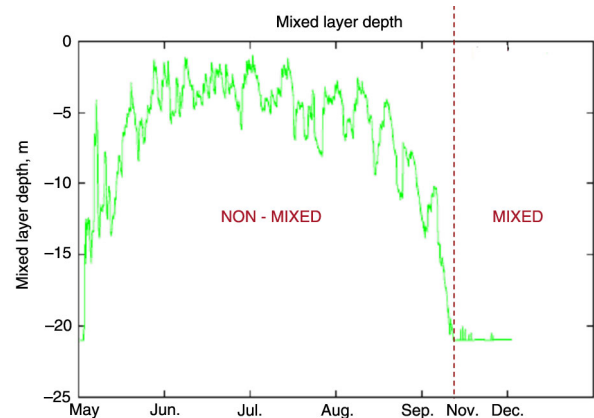

Fig. 1. Typical annual cycle of the mixed layer depth in a boreal (dimictic) 21 m deep lake as reproduced by FLake. The mixed regime takes place in autumn and early spring (for a short period), and the stratified regime takes place in summer. Diurnal oscillations, with the convective and wind-mixing regimes, are also represented.

- *Convective and wind-mixing regimes.* Typically, convection takes place during the night, while during the day mixing is mainly wind-generated. This results in diurnal oscillations of the mixed layer depth (h). In the convective regime h increases, and during wind mixing it decreases. These jumps are rapid and well reproduced by the lake model. So, on the diurnal scale the behaviour of h is also strongly non-linear, but at least all of the model variables are continuous.

In this study, DA is used to analyse the prognostic variables describing the liquid water reservoir; ice and snow variables are not analysed. Bottom sediments are not considered. The analysed variables are the mean water temperature ($\overline{T}$), the bottom temperature (T_b), the mixed layer depth (h) and the shape factor (C_T). Prognostic equations for the mixed and stratified regimes are described in Mironov (2008). The mixed layer temperature (T_{ML}) in FLake is calculated from a diagnostic equation (see Section 2.2).

The mean water temperature $\overline{T}$ and the lake bottom temperature T_b are slowly evolving variables. Depending on the depth of a lake and the season of a year, $\overline{T}$ varies on a time scale from several days to several weeks. For deep lakes, T_b changes gradually and shows only some annual variability. For shallow polymictic lakes (so shallow that their waters may mix from top to bottom several times throughout the open water period), oscillations of T_b during mixing periods follow oscillations of $\overline{T}$ and have a variability of several days. The variability of h strongly depends on the season. In summer, h usually shows significant daily oscillations. In early spring and autumn, during the mixing period, it is equal to the lake depth and remains constant. The integral of the polynomially approximated temperature profile in the thermocline C_T has a rather poor physical definition. During the stratified period, it usually changes gradually, with a time scale of several days. For the mixed regime, C_T does not exist mathematically, because there is no thermocline. In practice, in the model code, during the mixed regime C_T is kept constant and equals to a bogus value of 0.5.

Therefore, considering physical reasons and mathematical properties of the model, all lake water prognostic variables are controlled in the DA procedure. We considered separately (1) the ice and open water periods and (2) the mixed and stratified regimes for the open water period. Regarding daily oscillations of h during the stratified regime and the related model non-linearity, the time scale of these processes is equal to or shorter than the typical assimilation cycle period (1 d). Preliminary studies with perturbations of the initial state and analysis of the time evolution of different model variables showed less model sensitivity on a time scale of 1 d comparing to a time scale of several hours due to the smoothed diurnal oscillations of h. Therefore we did not distinguish the convective and wind-mixing regimes, and assumed that this non-linearity is acceptable for the assimilation procedure.

2.2. EKF application

By means of sequential DA, the control variables in the lake model should be initialised using information only from the LWST observations. In other words, innovations (departures between the observed and simulated LWST) should be redistributed among control lake variables. We pose the task differently for different lake model regimes. For the ice period, when there are no observations, the DA procedure is switched off. The assimilation cycling is stopped explicitly, because the gap between observations is too large (the entire winter), making it impossible to cycle the background error covariance matrix for this period. It is unknown beforehand, when observations will appear again after the ice break-up date, at mixed or stratified model regime. In an operational implementation, the question of skipping the analysis during the ice period should be specially addressed.

For the open water period and stratified regime, four prognostic variables are included into the state vector $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} \overline{T} & \eta & T_b & C_T \end{bmatrix}^{\mathrm{T}}.$$

Here $\overline{T}$ and T_b are the mean water and bottom temperatures (K), C_T is the shape factor, $\eta = 1 - \frac{h}{D}$ is the dimensionless mixed layer depth and h and D (both in m) are the mixed layer depth and the lake depth, respectively. For the mixed regime in FLake there is only one prognostic variable $\overline{T}$. The mixed layer depth h in this case is equal to the lake depth D, T_b is equal to $\overline{T}$ and C_T does not exist. Therefore, the state vector consists of only one variable:

$$\mathbf{X} = \begin{bmatrix} \overline{T} \end{bmatrix}.$$

We look for the analysis vector $\mathbf{X}_A$, using the previous forecast values for the background vector $\mathbf{X}_B$. Since only LWST is observed, and in open water conditions T_{ML} is equal to LWST, there is only one value of the observed mixed layer temperature T_{ML}^o in the observation vector $\mathbf{Y}$:

$$\mathbf{Y} = [T_{ML}^o].$$

In the case of the stratified regime the diagnostic equation for T_{ML} from FLake may be used for the observation operator $H(\mathbf{X})$:

$$T_{ML} = \frac{\overline{T} - C_T \eta T_b}{1 - C_T \eta}.$$

This expression and hence $H(\mathbf{X})$ is non-linear. In the case of the mixed regime, the observation operator $H(\mathbf{X})$ is

simply $T_{ML} = \overline{T}$. In both cases, a linearised observation operator $\mathbf{H}$ may be easily derived analytically. For the stratified regime, it can be given as:

$$\mathbf{H} = \left[\frac{1}{1 - C_T\eta} \ \frac{C_T(\overline{T} - T_b)}{(1 - C_T\eta)^2} \ \frac{-C_T\eta}{1 - C_T\eta} \ \frac{\eta(\overline{T} - T_b)}{(1 - C_T\eta)^2} \right].$$

For the mixed regime $\mathbf{H} = 1$.

Then, the EKF procedure may be applied in the form described in Le Moigne et al. (2009):

$$\mathbf{X}_B = M(\mathbf{X}_A)$$
$$\mathbf{B} = \mathbf{MAM}^T + \mathbf{Q}$$
$$\mathbf{K} = \mathbf{BH}^T (\mathbf{HBH}^T + \mathbf{R})^{-1}$$
$$\mathbf{X}_A = \mathbf{X}_B + \mathbf{K}(\mathbf{Y} - H(\mathbf{X}_B))$$
$$\mathbf{A} = (\mathbf{I} - \mathbf{KH})\mathbf{B}.$$

Here $M(\mathbf{X})$ is the non-linear model operator, $\mathbf{M}$ is the linearised model operator matrix, $\mathbf{B}$ is the background error covariance matrix, $\mathbf{R}$ is the observational error covariance matrix, $\mathbf{Q}$ is the model error covariance matrix, $\mathbf{A}$ is the analysis error covariance matrix and $\mathbf{K}$ is the Kalman gain vector. In our case, the observational error covariance matrix $\mathbf{R}$ is reduced to the scalar value of the observational error variance. For both the mixed and stratified regimes, the linearised model operator matrix $\mathbf{M}$ is obtained numerically by calculating Jacobians:

$$\mathbf{M} = \frac{\partial\, x_i^t}{\partial\, x_j^0}.$$

Here x_i^t are elements of the state vector; the lower index denotes the number of an element and the upper index denotes the time, from one sequential analysis cycle to another. For the stratified regime $\mathbf{M}$ is given as:

$$\mathbf{M} = \begin{bmatrix} \frac{\partial \overline{T}^t}{\partial \overline{T}^0} & \frac{\partial \overline{T}^t}{\partial \eta^0} & \frac{\partial \overline{T}^t}{\partial T_b^0} & \frac{\partial \overline{T}^t}{\partial C_T^0} \\ \frac{\partial \eta^t}{\partial \overline{T}^0} & \frac{\partial \eta^t}{\partial \eta^0} & \frac{\partial \eta^t}{\partial T_b^0} & \frac{\partial \eta^t}{\partial C_T^0} \\ \frac{\partial T_b^t}{\partial \overline{T}^0} & \frac{\partial T_b^t}{\partial \eta^0} & \frac{\partial T_b^t}{\partial T_b^0} & \frac{\partial T_b^t}{\partial C_T^0} \\ \frac{\partial C_T^t}{\partial \overline{T}^0} & \frac{\partial C_T^t}{\partial \eta^0} & \frac{\partial C_T^t}{\partial T_b^0} & \frac{\partial C_T^t}{\partial C_T^0} \end{bmatrix}.$$

In the mixed regime $\mathbf{M}$ is reduced to:

$$\mathbf{M} = \left[\frac{\partial \overline{T}^t}{\partial \overline{T}^0} \right].$$

To calculate Jacobians numerically, the state vector should be perturbed by values small enough for accurate approximation but not too small that round errors occur (Mahfouf et al., 2009). Our choice of perturbations resulted from a preliminary study with both small and large, negative and positive perturbations. We considered only the overall performance of EKF in terms of LWST evolution. For a better choice of perturbations, careful examination is

needed, as was done in Balsamo et al. (2004). In our study, for the stratified regime the following perturbation values of the state vector were used:

$$\delta\mathbf{X}^0 = [0.2K \ 0.05 \ 0.1K \ 0.05]^{\mathrm{T}}.$$

For the mixed regime, the state vector which has only one component, was perturbed by:

$$\delta\mathbf{X}^0 = [0.2K].$$

The size of matrix $\mathbf{B}$ changes when the regime changes. In these cases we stop the EKF cycling and reinitialise the $\mathbf{B}$ matrix. Explicit reinitialisation helps to avoid problems caused by non-linearity and possible EKF divergence (Jacobians and variances in the matrix $\mathbf{B}$ which are too large) when jumping between regimes. There is one more important constraint in the application of EKF in the FLake model. In the model initialisation, 'inverse' water density stratification is impossible. In terms of temperature, this means that in the stratified regime (1) the mixed layer temperature and the bottom temperature should both be higher or lower than the temperature of the maximum water density (277.13 K) and (2) if both of these are higher, the bottom temperature should be lower than the mixed layer temperature, and vice versa if both are lower. The crossover situation must not result from analysis. If it happens, the analysis is skipped and matrix $\mathbf{B}$ is reset.

In general, it is not trivial to specify the model error covariance matrix $\mathbf{Q}$, which contains model errors other than errors resulting from the initial state. In the case of offline simulations, when the lake model is driven by atmospheric forcing from a NWP model or from observations, it also contains errors from the forcing data. In our study, $\mathbf{Q}$ is prescribed in ad hoc manner. Variances were chosen as typical model errors squared. Typical model errors were derived from different studies, where the performance of FLake was assessed based on deep water temperature observations (Kirillin, 2010; Golosov et al., 2012). All correlations were prescribed to be 0.5, with signs determined using some modelling experience. For example, positive errors in the mean water temperature often lead to temperature profiles which are too sharp with mainly negative errors in the bottom temperature. Therefore, correlation between the appropriate errors is negative. Covariances are calculated accordingly. For better definitions of these values, more studies based on deep water temperature observations on lakes with different depths and mixing types are needed. For the mixed regime, $\mathbf{Q}$ is defined as

$$\mathbf{Q} = [4.0K^2],$$

and for the stratified regime as

$$\mathbf{Q} = \begin{bmatrix} 4.000K^2 & 0.100K & -1.000K^2 & -0.100K \\ 0.100K & 0.010 & -0.050K & 0.005 \\ -1.000K^2 & -0.050K & 1.000K^2 & 0.050K \\ -0.100K & 0.005 & 0.050K & 0.010 \end{bmatrix}.$$

In our experiments, matrix $\mathbf{B}$ evolved from matrix $\mathbf{Q}$ in the first cycle, and was reset back to $\mathbf{B}=\mathbf{Q}$ each time at reinitialisation. The sensitivity of the analysis system performance to $\mathbf{Q}$ values was studied only in terms of LWST and appeared to be small (0.1°C order of magnitude).

3. Lake observations

A good overview of lake observations, in-situ and remote-sensing, is given in Kheyrollah Pour et al. (2014a, 2014b). Both types of observations have advantages and disadvantages. Space-borne observations not only contain information about LWST and ice cover, but also about other lake water characteristics, for example, water quality parameters (Potes et al., 2011). LWST and ice cover are measured by different sensors on various satellites. In Kheyrollah Pour et al. (2014b), the MODIS data are mainly used. Different satellite observations for 263 large lakes from the ARCLake project (http://www.geos.ed.ac.uk/arclake/) (MacCallum and Merchant, 2012) were compared with operational LWST analyses in the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) product by Fiedler et al. (2013). However, remote-sensing data may contain gaps due to cloudiness. Errors due to the fractional ice cover, undetected clouds and inaccuracies in the land-water mask may be significant. Also, the remotely measured water skin temperature may differ from the bulk temperature, which is predicted by models and measured in-situ [see, e.g., Donlon et al., (2002)].

In-situ measurements may not only contain information about lake surface characteristics but also information about the vertical temperature distribution, ice depth and other properties [see, e.g., Arst et al. (2008) and Jonas et al. (2003)]. However, in-situ data are produced only for selected lakes and are used mainly for research purposes, with the exception of operational SYKE measurements. For operational in-situ measurements, the time period is problematic. Usually instruments are installed in early spring when the ice has already disappeared. This happens approximately 1 week after the ice break-up and means that these measurements alone do not provide the exact ice break-up date. Therefore, extra observations are needed. At present, both types of lake observations are widely used for model verification (Duguay et al., 2003; Kheyrollah Pour et al., 2012). In the NWP context, lake state observations are used in a very simplified way (Eerola et al., 2010; Rontu et al., 2012; Kheyrollah Pour et al., 2014b).

In our study, two sets of LWST observations were assimilated into the lake model. First, we assimilated in-situ SYKE observations for 27 lakes in Finland (see Table 1 for the lake locations and depth). SYKE observations do not contain direct information about ice cover, and during the ice period LWST is not measured. Therefore, the 'no data' flag indicates that LWST is not observed either because of ice cover or for other reasons. In practice, the majority of cases with the 'no data' flag correspond to the ice cover situation, and these periods last for several months. In our experiments, for cases with 'no data' flag and during the ice period, the assimilation cycling was

Table 1. The impact I (for the definition see Section 4.2) of the assimilation of SYKE observations of LWST for the open water period (EKF-S experiment, summer of 2011) for 27 lakes whose geographical coordinates (deg) and mean depth D are given

Name (longitude, latitude)	D (m)	I (%)	Name (longitude, latitude)	D (m)	I (%)
Kuivajärvi (23.9, 60.8)	2.2	94.8	Rehja-Nuasjärvi (28.0, 64.2)	8.5	95.5
Tuusulanjärvi (25.1, 60.4)	3.2	94.3	Vaskivesi (23.8, 62.1)	7.0	97.1
Pääjärvi 1 (24.5, 62.9)	3.8	96.6	Haukivesi (28.4, 62.1)	9.1	94.9
Pesiöjärvi (28.7, 64.9)	3.9	95.4	Kallavesi (27.7, 62.8)	9.7	96.3
Kyyvesi (27.1, 62.0)	4.4	96.5	Pielinen (29.6, 63.3)	10.1	94.6
Jääsjärvi (26.1, 61.6)	4.6	96.2	Konnevesi (26.6, 62.6)	10.6	95.4
Nilakka (26.5, 63.1)	4.9	96.6	*Saimaa* (28.1, 61.3)	10.8	94.5
Pyhäjärvi (22.3, 61.0)	5.5	96.4	Ala-Rieveli (26.2, 61.3)	11.2	92.4
Längelmävesi (24.4, 61.5)	6.8	94.4	Päijänne (25.5, 61.6)	14.1	93.7
Ounasjärvi (23.6, 68.4)	6.6	97.3	*Inarijärvi* (27.9, 69.1)	14.3	97.1
Lappajärvi (23.7, 63.1)	6.9	93.4	Näsijärvi (23.8, 61.6)	14.7	94.0
Oulujärvi (27.0, 64.5)	7.0	95.0	Pääjärvi 2 (25.1, 61.1)	14.8	96.7
Unari (25.7, 67.1)	7.0	94.0	Kilpisjärvi (20.8, 69.0)	19.5	96.8
Kevojärvi (27.0, 69.8)	7.0	98.0			

interrupted and the model evolved without any analysis update (relying on the model state). When LWST measurements are again available, the background error matrix was reinitialised and the assimilation cycling was resumed from the current state vector.

Secondly, for four selected lakes in Finland, in-situ measurements from SYKE were combined with MODIS UW-L3 data prepared by the University of Waterloo (Kheyrollah Pour et al., 2014a) for early spring, and observations from SYKE about ice break-up dates [see Kheyrollah Pour et al. (2014b) for details]. The four lakes differ in depth, size and location: (1) deep northern Lake Inarijärvi, (2) shallow southern Lake Tuusulanjärvi, (3) large Lake Saimaa and (4) medium Lake Lappajärvi (see Table 1 for details). Time series of the LWST measured by SYKE and observed by MODIS were merged manually. MODIS data were taken for the pixels nearest to the SYKE locations and added to the time series starting from the ice break-up date until the first spring SYKE measurements date. For the period prior to the ice break-up date, the 'no data' flag was set to LWST, as in the original SYKE measurements. Hereafter, we refer to this type of data as 'the merged data set'. The main aim of the experiments with the merged data set was to study the role of observations during early spring. Cross-validations were also performed with these data.

LWST is measured daily at 08.00 Eastern European Time (EET) by SYKE. One MODIS observation per day was added to the merged data set. In our experiments, all MODIS data were assigned the time 08.00 EET (note that there is almost no diurnal cycle of LWST in the early spring). Gaps in MODIS observations were filled manually with time-interpolated values. Assuming in-situ measurements provided by SYKE were of good quality and manually selected MODIS observations were carefully checked, no other quality controls were used.

The observational error variance usually consists of the instrumental error squared and the representativeness error variance. For MODIS, the error variance associated with the skin effects, undetected cloudiness and floating ice should be included. In all of our experiments, the observational error variance is equal to an ad hoc value of $1.0\,\text{K}^2$. This is the order of magnitude of the difference between the in-situ and MODIS observations variance (Kheyrollah Pour et al., 2012). The specification of observational error statistics requires further studies. Note that the observational error variance is small compared to the background error variances (which represent the errors of the parameters defining the lake water temperature profile and may be compared with the observational error variance using values of the **H** vector, see Appendix).

4. Numerical experiments

4.1. Experiment setup

The EKF algorithm for assimilating LWST observations was implemented in a stand-alone (1D) offline version of the lake model FLake. In the offline mode the lake model is coupled only one-way with a NWP model, being forced by atmospheric model data. Forcing data from operational runs of the HIRLAM NWP model were used. The following forcing variables were extracted from the HIRLAM 7.3 operational 6-hour forecasts[1] archives: wind speed, temperature and specific humidity at the lowest model level, accumulated downward long-wave and short-wave radiation fluxes and the total snowfall. Hourly outputs were used to calculate radiation fluxes and precipitation rates. The turbulent heat and momentum fluxes were calculated with the specific flux formulation in FLake.

Experiments were performed for a 1 yr period from 3 November 2010 to 10 November 2011. For this period, observations from different lakes were selected and compiled into time series. For both the SYKE and merged data sets, observations were assigned to the valid time 08.00 EET. The EKF analysis was performed at 06.00 EET with a cycling period of 24 hours and an assimilation window[2] of 4 hours (we considered all observations 2 hours before and 2 hours after the analysis time and referred them to this time). Note that in our experiments the satellite observations were put into the assimilation window artificially. Therefore, in operational implementation, the assimilation window question should be specially addressed. For cross-validation, observations were assimilated from every second day (the rest were used for validation). In this case, the cycling period was 48 hours. For all lakes, the free model run (without DA) was initialised from the typical late autumn temperature profile for a boreal lake: a lake is mixed down to the bottom, and the water temperature is equal to the screen level temperature. Numerical experiments with FLake-EKF were performed for 27 lakes assimilating SYKE observations and for four lakes assimilating merged observations. The design of experiments is summarised in Table 2.

---

[1]The model horizontal resolution is 15 km, 60 levels in vertical.

[2]The following notation is used here: An analysis time is a moment in time when the analysis is performed to initialise the new forecast (increments are added to the background). A cycling period is a period between two analysis times. The (shortest) forecast length equals to the cycling period. Jacobians are calculated with the model runs for the full previous cycling period. Assimilation window is a time period around the analysis time, during which the available observations are picked to be assimilated.

The best way to evaluate the performance of the assimilating system is by a comparison with independent observations. Either different sources of data may be used, or cross-validation methods may be applied, when observations from the single source of data are partly assimilated, partly used for validation. A posteriori verification is useful to reveal potential problems, for example, biases in a model or in observations. The quality of the linear approximation and possibilities for simplifications (if any) may be assessed from the behaviour of the Jacobians, as well as from the evolution of the **B** matrix and Kalman gain vector **K**. However, to study in details the evolution of all the EKF components, the amount of material is large. The statistics differ depending on the season and the corresponding model regime. In this paper, the entire picture is described, focusing on the changing regimes and other physical aspects. Only a preliminary study of Jacobians, the components of the **B** matrix and vector **K** is performed. In experiments using different datasets we evaluate the performance of the EKF depending on data availability. For SYKE observations (EKF-S experiment) the impact of data is shown for all of the lakes and the behaviour of the components of the matrices and vectors **M**, **H**, **B** and **K** is briefly described. For the merged data set (EKF-M experiment), the results of cross-validation are presented. A detailed study of a posteriori statistics and the performance of the EKF assimilation system in terms of the evolution of its components is planned for the future.

4.2. Results of experiments and discussion

4.2.1. Assimilation of SYKE data. A number of examples are given to provide a qualitative overview of the assimilation of SYKE observations. In the FR experiment (the free run) T_{ML} was too warm for all of the lakes in summer. This is a well-known FLake problem (Kourzeneva et al., 2012b; Martynov et al., 2012). Figure 2 shows that for Lake Inarijärvi the mixed layer temperature T_{ML} in the EKF-S experiment is much closer to observations than in the FR experiment, where it is up to 5°C warmer. In Fig. 3, comparing the EKF-S experiment with the FR experiment, no

Table 2. Design of the experiments

Experiment name	Data assimilation	Observations	Cycling period	Number of lakes
FR (free run)	No	No	No	27
EKF_S	Yes	SYKE	24 h	27
EKF_M	Yes	Merged	24 h (48 h for cross-validation)	4

impact is seen on the ice and snow thicknesses in autumn and winter (the snow thicknesses in EKF-S and EKF-M are not shown), mainly because of the lack of SYKE observations just before the ice onset date. In late autumn and winter, the EKF-S experiment relied on a background state similar to the FR experiment. In spring, the FR experiment usually showed an ice break-up date which was too late. The early spring situation for different experiments is described in Table 3. For Lake Inarijärvi in the FR experiment the ice break-up was on June 6th. For the EKF-S experiment it was 2 weeks earlier, on May 23rd, and coincided with the beginning of SYKE observations in spring. However, the SYKE observations usually commence one or 2 weeks after the actual ice break-up date. Therefore, LWST may appear to be much higher than the melting point. For example, in Table 3 in the EKF-S experiment the observations for Lake Inarijärvi start from the already high LWST value of 7.6°C. Thus, the real ice break-up conditions were unknown from both the FR and EKF-S experiments. In Fig. 2, the late ice break-up date in the FR experiment influences the spring T_{ML} (values too low), so that T_{ML} in June is up to 10°C lower for the FR experiment than for the EKF-S experiment. The EKF-S experiment compared with FR gave much warmer T_{ML} during the whole of spring. Hence, with SYKE observations alone (without satellite observations), there was lot of uncertainty in real knowledge of the lake state, mainly in early spring but also in late autumn. The impact of the EKF assimilation on the ice and snow temperatures in autumn and winter was minor (not shown).

Through the DA, the LWST observations also influenced the mean water temperature $\overline{T}$ and the bottom temperature T_b (Fig. 4 for Lake Inarijärvi and Fig. 5 for Lake Saimaa). The impact was large, and typically the summer maximum was warmer in the EKF-S experiment than in FR: for 5–7°C for $\overline{T}$ and for 7–11°C for T_b. Since deep water temperature measurements were not available, it is difficult to understand which experiment is closer to the truth. For a better understanding, if T_b in the EKF-S experiment is realistic, it was compared with the observed T_b values in Lake Valkea-Kotinen (Dmitrii Mironov, personal communications 2014). Lake Valkea-Kotinen is a very small lake with the mean depth of 3 m. It is located in Southern Finland, about 100 km to the northwest of Lake Tuusulanjärvi. Both lakes are small and shallow and located in the same climate zone, and are expected to exhibit similar behaviour annually. The maximum summer T_b of Lake Valkea-Kotinen was 7°C in 2006. For Lake Tuusulanjärvi T_b was 20.5°C in the EKF-S experiment (not shown), which is too high. For Lake Saimaa T_b shown in Fig. 5b is 18.2°C, which is also too high. According to preliminary studies, the sensitivity of $\overline{T}$ and T_b evolution to the **Q** matrix specification is high (not shown).

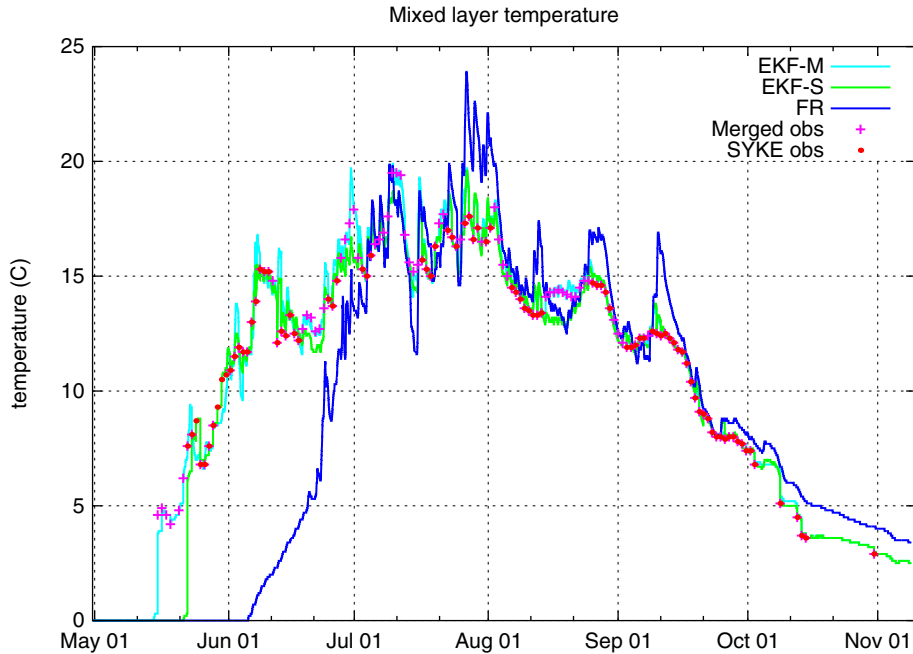*Fig. 2.* Time evolution of the mixed layer temperature in °C for Lake Inarijärvi (the mean depth is 14 m) for the summer period from May 2011 to November 2011. The FR, EKF-S and EKF-M results are shown by the blue, green and cyan lines, respectively. The SYKE and merged LWST observations are represented by the red dots and pink crosses, respectively.
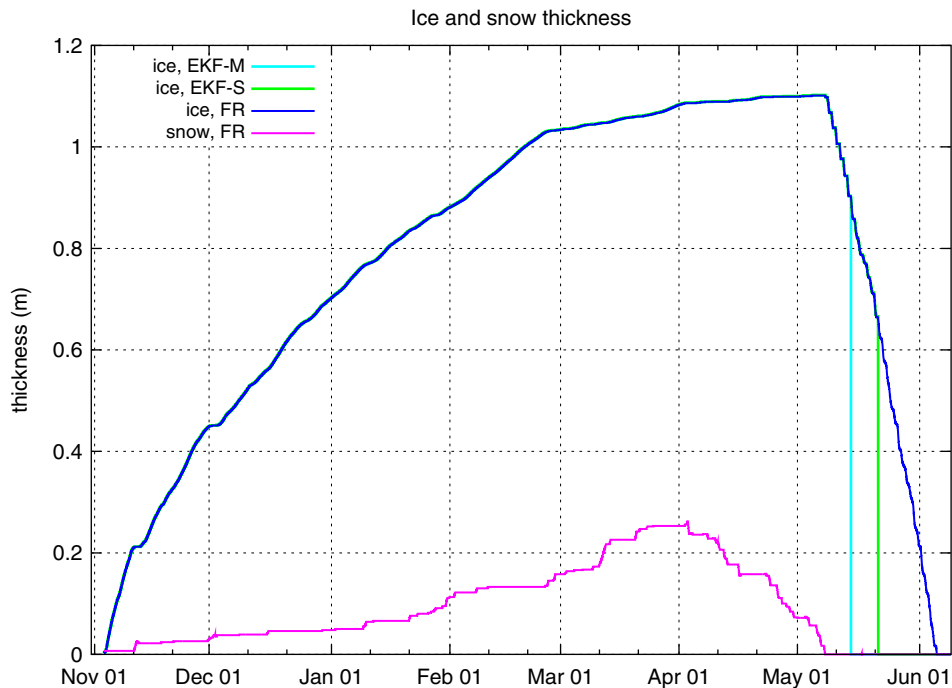

Fig. 3. Time evolution of the ice and snow thickness in metres for Lake Inarijärvi (the mean depth is 14 m) for the winter–spring period from November 2010 to June 2011. The FR, EKF-S and EKF-M results for ice thickness are shown by the blue, green and cyan lines, respectively. The FR results for snow thickness are shown in pink. The EKF-S and EKF-M results for the snow thickness coincide with the FR results and are not shown.

Table 3. Ice break-up dates in the FR, EKF-S and EKF-M experiments and the first spring observed LWST for the EKF-S and EKF-M experiments (note that the ice break-up dates are contemporaneous with the spring start of observations for EKF-S and EKF-M)

		Inarijärvi	Saimaa	Lappajärvi	Tuusunlajärvi
FR	Date	06/06	05/10	05/09	05/07
EKF-S	Date	05/23	05/11	05/13	05/06
	LWST	7.6°C	12.6°C	12.0°C	8.9°C
EKF-M	Date	05/16	05/05	05/04	04/26
	LWST	4.6°C	7.8°C	2.7°C	2.8°C

Thus, poorly specified values in the **Q** matrix may be one reason for the overestimation of the deep water temperatures. Better specification needs further study, comparing the modelling results with regular deep water temperature observations. Another possible reason is the wrong timing of regimes. Both T_b and $\overline{T}$ evolve slowly, depending on the lake depth. In the EKF-S experiment, T_b and $\overline{T}$ were strongly influenced by memory from the early spring just after ice break up, when a lake is mixed down to its bottom. Usually this mixed period is rather short (several days). In most cases, SYKE observations started only after this period. At that moment, the lake model still showed a mixed regime, while in reality the regime was already stratified. The EKF algorithm resulted in positive corrections to the entire water column (both for $\overline{T}$ and T_b) instead of making a non-mixed profile. Due to the long memory of T_b and $\overline{T}$, this resulted in their overestimation for the entire modelling period. In this way, the lack of SYKE observations in the early spring maintained errors of 6–11°C in deep water temperatures during the entire summer.

Although the state vector includes the dimensionless mixed layer depth η, the mixed layer depth h itself (connected to η by the simple equation $h = D(1 - \eta)$, where D is the lake depth) can be interpreted more naturally. In contrast with deep water temperatures, h is a quickly varying variable. It was also influenced by the EKF analysis. In Fig. 6a, the EKF-S experiment shows earlier autumn mixing than FR. In summer, the EKF-S experiment usually showed a deeper mixed layer. But since h changes quickly, it tends to return to its background state within one DA cycle, which neglects the assimilation impact. Results for Lake Saimaa (11 m deep) illustrate this summer process in more detail (Figs. 7–9). In Fig. 7, the h increments are mainly negative and the $\overline{T}$ increments are very small. Thus, the EKF-S experiment produced colder T_{ML} values, which were closer to the observations, but the result was unstable (Fig. 8). As h quickly returned to its background state (i.e. wrong values), T_{ML} increased quite fast and again deviated from the observations. In Fig. 9b the innovations are mainly negative. Physically, in this situation the model jumps between the convective and wind-mixing regimes. During daytime in summer, due to short-wave radiation,

the temperature of the upper-most water layer increases and stratification becomes stable (the wind mixing regime with the absolute value of h decreasing). During night-time, due to long-wave cooling, the temperature of the upper-most surface layer decreases and stratification becomes unstable (the convective regime with the absolute value of h increasing). In the model, these two regimes are described by different equations. The mainly negative increments of h in the EKF analysis mean that the prognostic equation for h in the stable regime may be biased. The problem is enhanced by $\overline{T}$ in summer which is too warm in EKF-S. During summer, when the absolute value of h is small, the deep and upper water layers are decoupled. This results in small $\overline{T}$ analysis increments (Fig. 7b), not enough to fix existing errors (see also the analysis of the tangent linear model operator **M** later in this section). Assimilation of more observations, for example, from remote sensing, more frequent and starting earlier in the spring, might improve the situation. A possible bias in the atmospheric forcing may also be a reason for the discrepancy: absolute values of h which are too small may come from momentum fluxes being too small. It is not easy to answer the question, whether excluding h from the state vector would be a good solution of the decoupling problem: it only evolves quickly in midsummer, but more slowly in spring and autumn. Besides, it is responsible for the timing of the mixed and stratified regimes. Figure 8 illustrates also that in the EKF experiments the observational error variance is small compared to the background error variances: the EKF-S results are very close to the observations.

The shape factor C_T has no clear physical meaning, thus it is difficult to analyse its performance. It can be said that it was also influenced, and in Fig. 6b there is a significant difference between the FR and EKF-S experiments.

In order to quantify the assimilation performance with a posteriori statistics, a root mean square error (RMSE) was computed for different lakes. The RMSE was calculated for the open water period, when observations of LWST were available. The impact I (%) of the assimilation with respect to the model was calculated as:

$$I = \frac{RMSE_{\text{mod}} - RMSE_{assim}}{RMSE_{\text{mod}}} \cdot 100\%$$
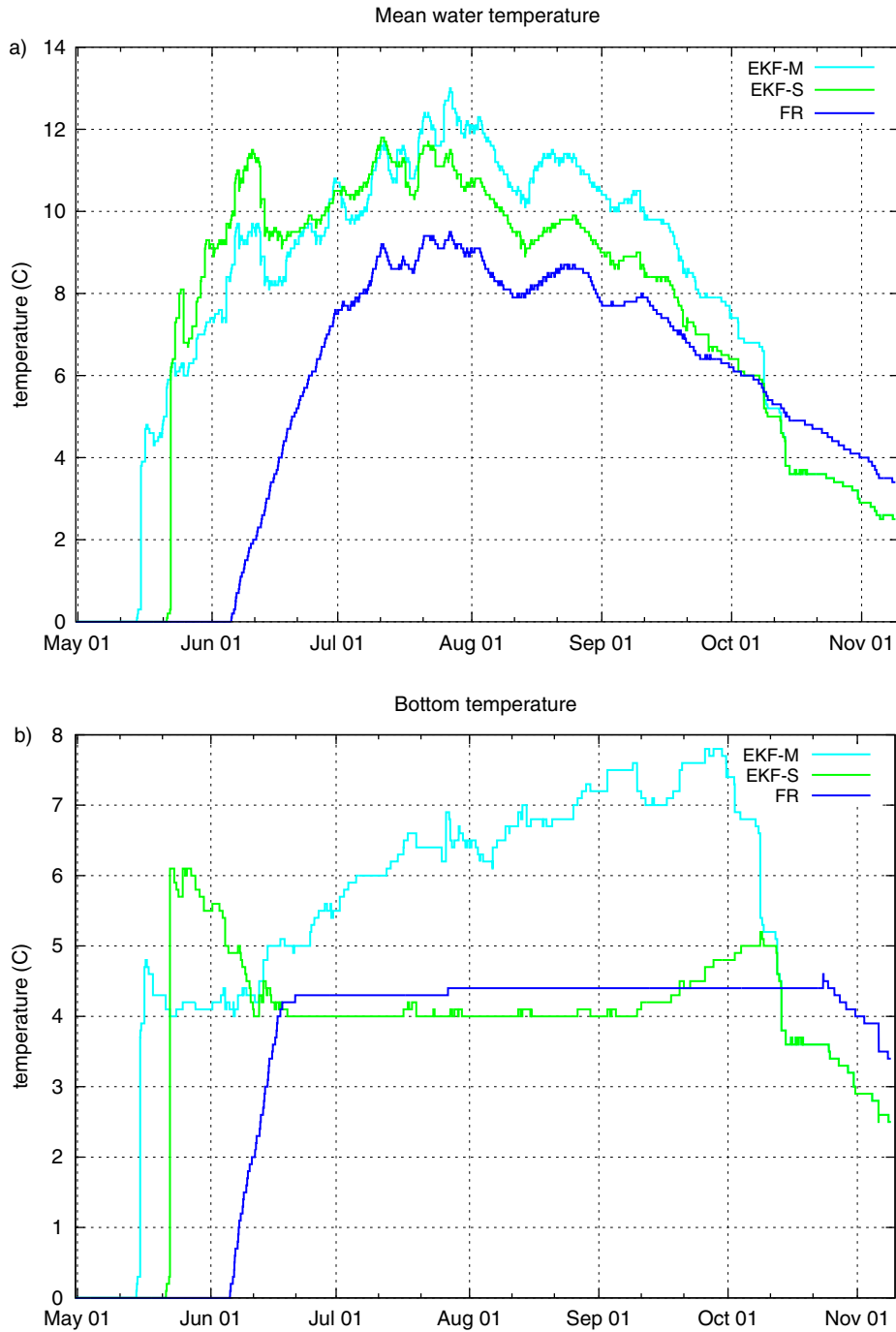
Fig. 4. Time evolution of (a) the mean water temperature and (b) the bottom temperature in °C for Lake Inarijärvi (the mean depth is 14 m) for the summer period from May 2011 to November 2011. The FR, EKF-S and EKF-M results are shown by the blue, green and cyan lines, respectively.

Here $RMSE_{\mathrm{mod}}$ and $RMSE_{assim}$ are for the free model run and for the assimilation system run, respectively. The impact for different lakes along with their geographical coordinates and mean depths is presented in Table 1. As expected, $RMSE_{assim}$ is much smaller than $RMSE_{\mathrm{mod}}$ and

the impact is high (note that the analysed LWST is not a weighted average between the observed and modelled values). A maximum impact of 98.0% is seen for the medium-depth polar Lake Kevojärvi and a minimum impact of 93.4% for the medium-depth Lake Lappajärvi
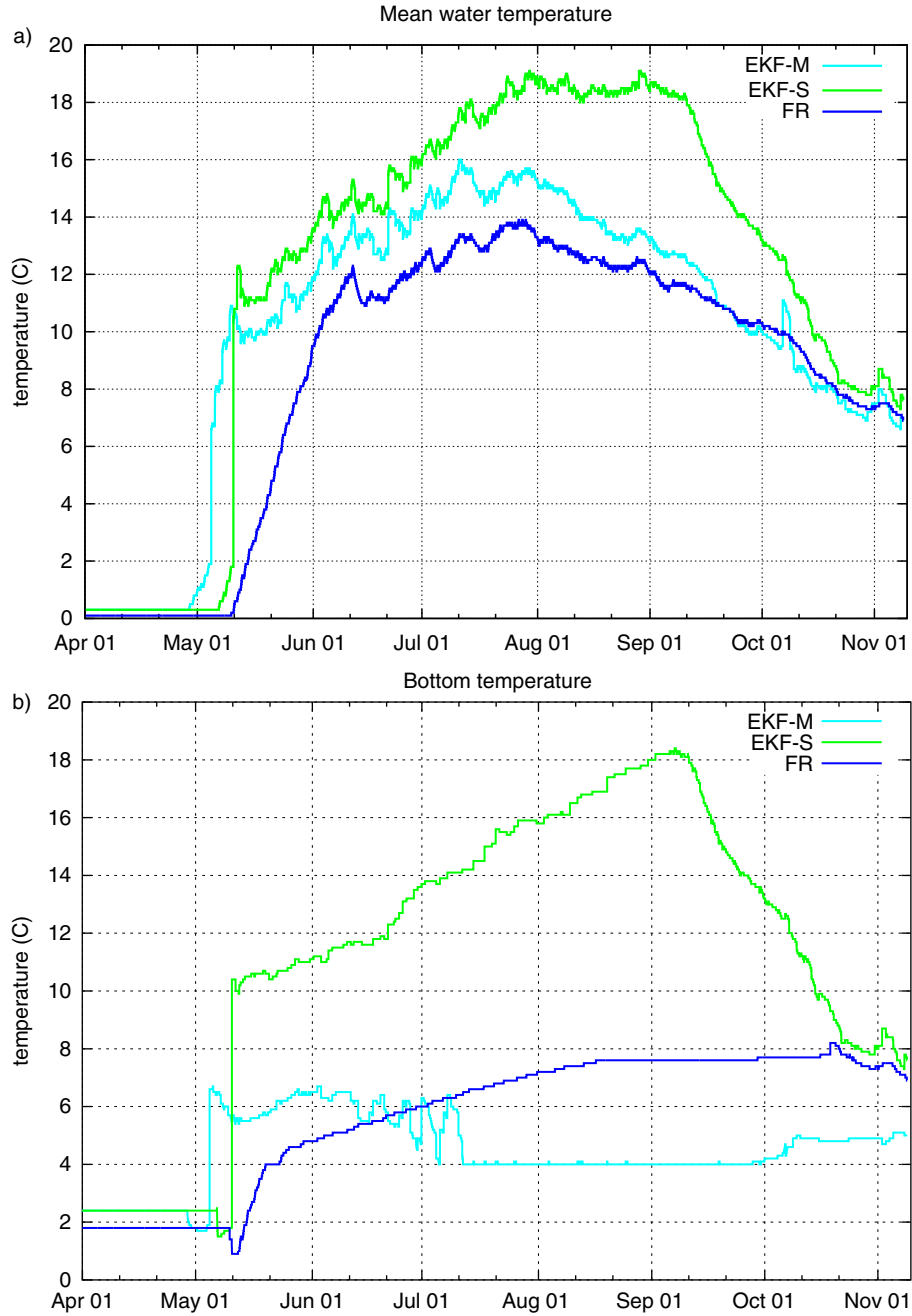
Fig. 5. Time evolution of (a) the mean water temperature and (b) the bottom temperature in °C for Lake Saimaa (the mean depth is 11 m) for the summer period from April 2011 to November 2011. The FR, EKF-S and EKF-M results are shown by the blue, green and cyan lines, respectively.

located in central Finland. A dependency of the impact on the lake depth or latitude is not seen (the lakes in Table 1 are arranged by increasing mean depth).

Weights in the Kalman gain vector **K** define the overall performance of the analysis system and depend on the components of the background errors covariance matrix **B**. Behaviour of the **B** matrix is in turn dependent on the

components of the tangent linear model operator **M** and on the linearised observation operator **H** (note that the components of **H** are derived analytically). The detailed study of their behaviour is planned for the future. A preliminary study was performed with SYKE observations for Lake Inarijärvi and Lake Tuusuljärvi, perturbation values and the matrix **Q** indicated in Section 2. In the Appendix
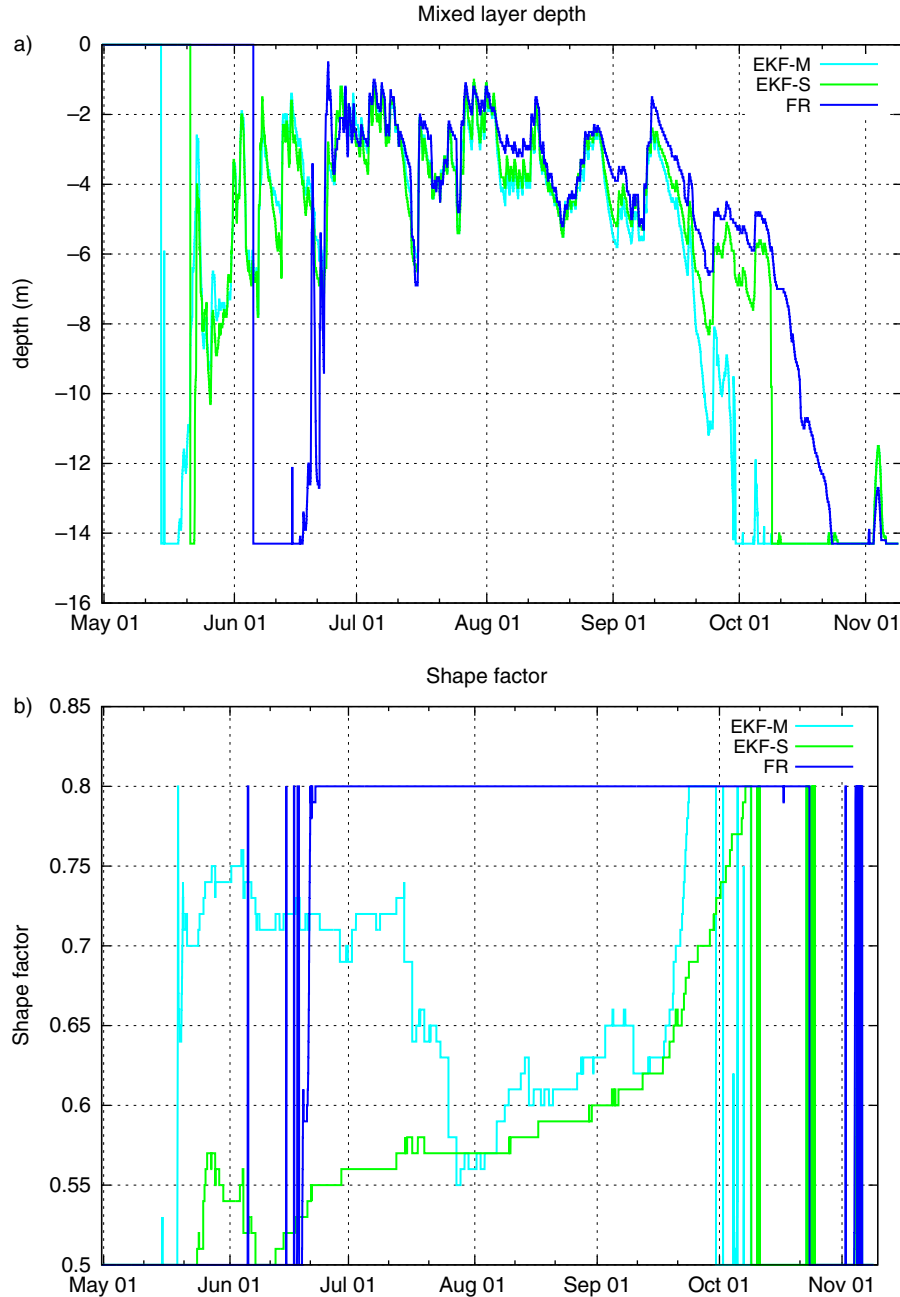
Fig. 6. Time evolution of (a) the mixed layer depth in metres and (b) the shape factor for Lake Inarijärvi (the mean depth is 14 m) for the summer period from May 2011 to November 2011. The FR, EKF-S and EKF-M results are shown by the blue, green and cyan lines, respectively. Note that the shape factor changes rapidly and during mixing period it equals to a bogus value of 0.5.

the limits of the components of the matrix **M**, the vector **H**, the matrix **B** and the vector **K** are shown and their behaviour at different mixing regimes is briefly described. For the stratified regime, many components of the matrix **M** and the vector **H**, as well as of the matrix **B** and the vector **K** vary depending on season and lake type and often show an annual cycle. Maxima and minima appear in mid-summer when stable stratification prevails, or in autumn

before mixing occurs. The components of the matrix **M** are bounded and evolve smoothly, which means that there is a good tangent linear approximation within each model regime. Sometimes the components of the matrix **M** and vector **H** can be very large (e.g. $\frac{\partial T_b^t}{\partial C_T^0}$, $\frac{\partial T_{ML}}{\partial \eta}$, $\frac{\partial T_{ML}}{\partial C_T}$). This leads to large values in the matrix **B** elements (e.g. var (e_{T_b})). But being large, they remain bounded, and decrease when the appropriate components of the **M** matrix and the vector **H**
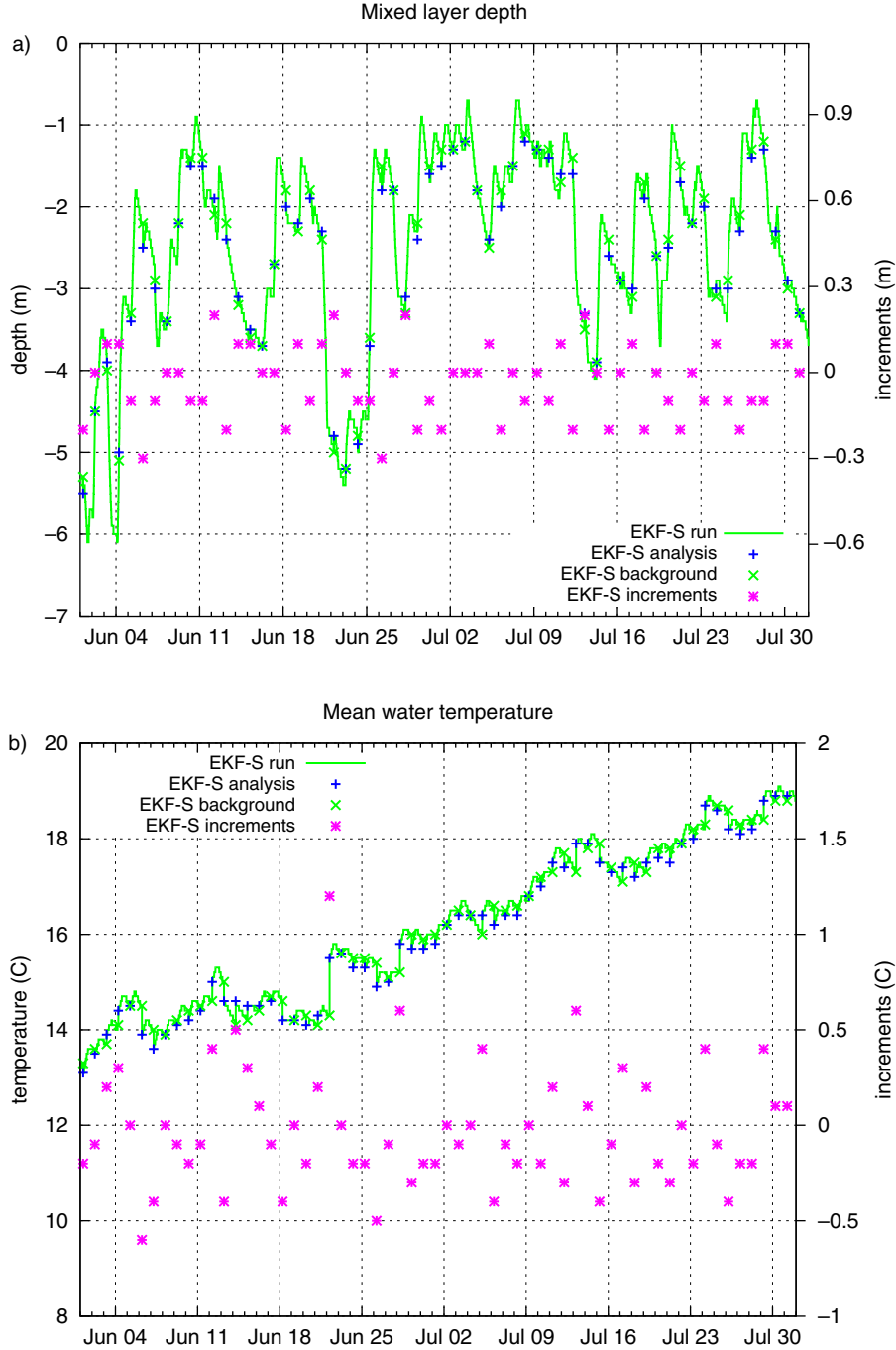
Fig. 7.   Time evolution of increments, background and analysis values of (a) the mixed layer depth in metres and (b) the mean water temperature in °C for Lake Saimaa (the mean depth is 11 m) for the period of June–July 2011. The EKF-S results are shown by the green line; the analysis increments, the background and analysis values are represented by the pink, green and blue crosses, respectively.

decrease (as well as when the **B** matrix is reset). Some components of the matrix **M**, such as $\frac{\partial T'_h}{\partial T^0_h}$, remain almost constant and are the same for different lakes. Components of the Kalman gain vector **K** show an annual cycle or oscillations in time. All of these are stable, even when some of the **B** matrix components are large. For the mixed

regime, all of the matrices and vectors are reduced to scalar values and remain almost constant in time (although $\frac{\partial \overline{T}'}{\partial \overline{T}'}$ is slightly less for shallow lakes). The almost constant values give potential for simplification of the EKF algorithm. Note that in midsummer $\frac{\partial \overline{T}'}{\partial \overline{T}^0}$ has minimum and $\frac{\partial \overline{T}'}{\partial \eta^0}$ has maximum. This corresponds to decoupling and leads to
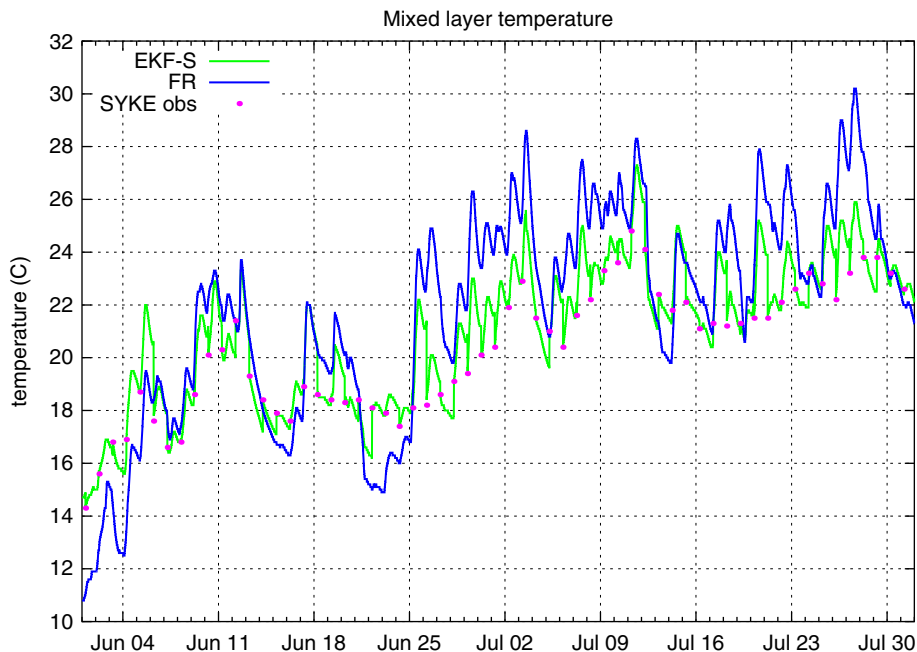
Fig. 8. Time evolution of the mixed layer temperature in °C for Lake Saimaa (the mean depth is 11 m) for the period of June–July 2011. The FR and EKF-S results are shown by the blue and green lines, respectively. The LWST observations are represented by the pink dots.

small weights in the Kalman gain vector **K** and small analysis increments in $\overline{T}$. Increasing the related components of the **B** matrix in order to propagate information in the vertical in the decoupling situation and to increase the weights for $\overline{T}$, necessitates modification of the filtering procedure.

4.2.2. Assimilation of merged data and cross-validation. Experiments with the merged data set were designed to study the role of LWST information in early spring, just after ice break-up. Can this information improve the quality of the analysis? To answer this question, numerical experiments with two observational datasets, SYKE and merged (EKF-S and EKF-M experiments), were compared. The ice break-up dates in the EKF-S and EKF-M experiments did not differ much. In the EKF-M experiment, the ice break-up date was up to 7 d earlier; however, the first spring LWST was up to 9°C higher than observed (see Figs. 2 and 3, and Table 3). Thus, T_{ML} in the EKF-M experiment increased very quickly during this 7-d period. In Fig. 2, the T_{ML} difference between the EKF-M and EKF-S experiments reaches 10°C in early spring, and then diminishes within 2–3 weeks. Therefore, starting from June the effect of additional early spring observations on T_{ML} cannot be noticed. Spring mixing in the EKF-M experiment appeared also several days earlier (Fig. 6a). The summer behaviour of the temperature profile in lakes was also influenced by early spring observations. The impact

was different depending on the spring mixing conditions. For Lake Saimaa, the spring mixing period lasted for 2–3 d (not shown). For this lake, the summer maxima of $\overline{T}$ and T_b in the EKF-M experiment were up to 4°C and 12°C lower compared with the overestimated values in the EKF-S experiment (Fig. 5) and therefore seem to be more realistic. For the very shallow Lake Tuusulanjärvi (the mean depth is 3 m) the spring mixing period lasted for 15–20 d, and in terms of $\overline{T}$ and T_b there was no difference between the EKF-S and EKF-M experiments, although in the EKF-M experiment the spring T_{ML} was higher until June (not shown). Different summer behaviour of the lake temperature profiles in the EKF-M and EKF-S experiments resulted in different autumn mixing conditions. For Lake Inarijärvi, there was earlier mixing in the EKF-M experiment (Fig. 6a). This in turn may influence the ice onset date the following year and the following winter ice conditions. The length of our experiments was not sufficient for studying these aspects and this should be investigated in future.

In the EKF-M experiments, there was also a problem of a possible bias in the prognostic equation for h in the case of stable stratification. Improved summer values of $\overline{T}$ resulted in neither more stable behaviour of T_{ML} nor smaller increments of h. In summer observational departures of T_{ML} were mainly negative and increments of h were mainly positive (not shown) and had approximately the same values as in the EKF-S experiments. This means that the problem of a possible bias in the prognostic equation for h cannot be corrected by better $\overline{T}$ values due to
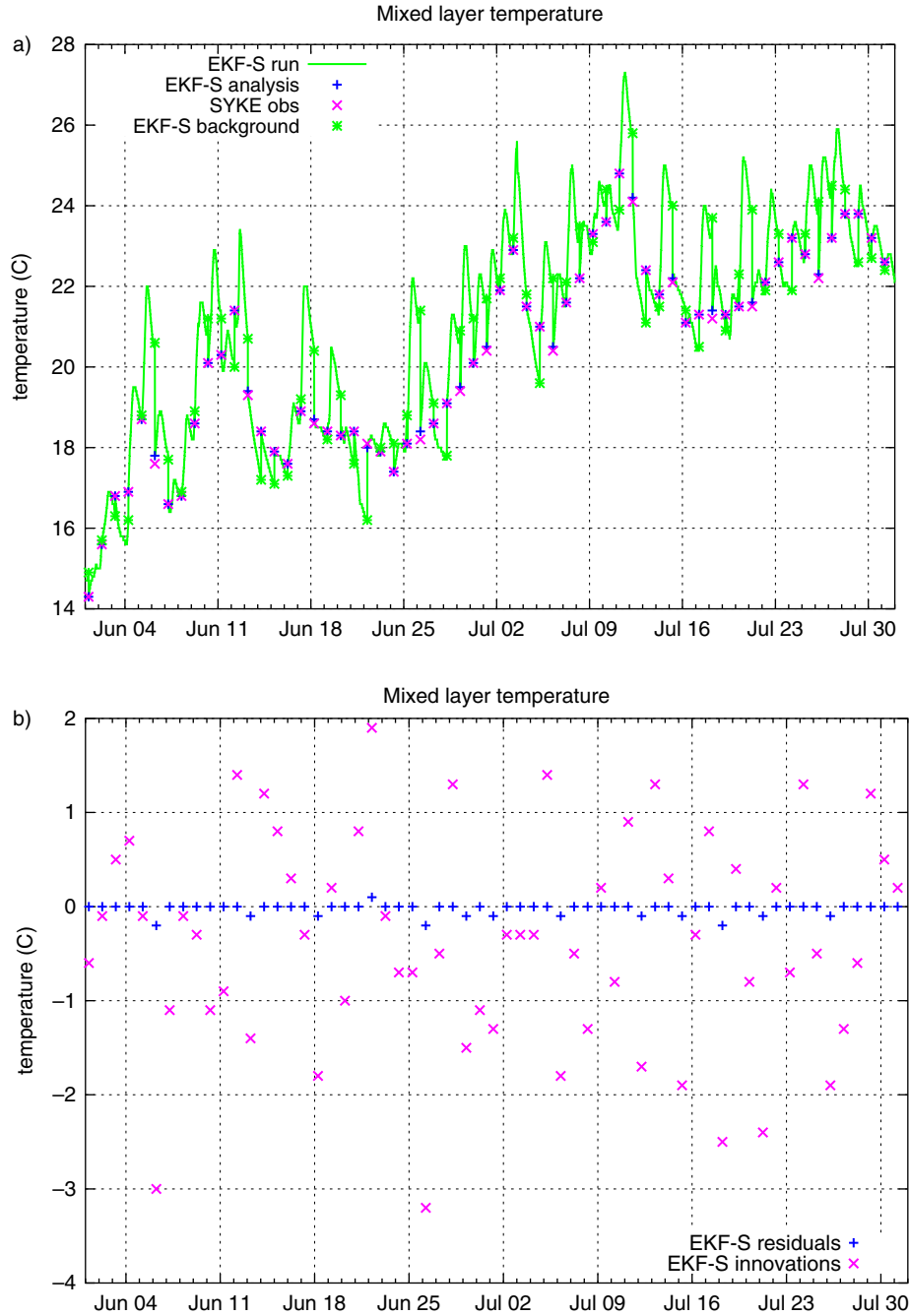
Mixed layer temperature

Fig. 9. Time evolution of (a) the mixed layer temperature and (b) the innovations (observed minus background values) and residuals (observed minus analysed values) in °C for Lake Saimaa (the mean depth is 11 m) for the period of June–July 2011. The EKF-S results are shown by the green line. The LWST observations, background and analysis values are represented by the pink, green and blue crosses, respectively. The innovations and residuals are represented by the pink and blue crosses, respectively.

additional early spring observations and necessitates spe-cial attention. Perhaps, bias correction in the assimilation system should be considered.

For the merged data set, the assimilation system was cross-validated, assimilating every second observation and using the other data to calculate scores. Note here a longer cycling period (48 hours) was used, which means more non-linearity in the model, less accurate Jacobians and larger analysis departures (not shown). The results of the cross-validation, as well as the impact for the merged data, are presented in Table 4. Here, the bias, RMSE and the error standard deviation (ESTD) were calculated for each lake

Table 4. The impact *I* (for the definition see Section 4.2), the bias, the root mean square error (RMSE) and the error standard deviation (ESTD) (°C) for the assimilation of the merged LWST observations for different lakes (the bias is calculated as simulated minus observed values, a positive bias means an overestimation of LWST)

Lake	*I* (%)	FR			EKF-M		
		Bias	RMSE	ESTD	Bias	RMSE	ESTD
Inarijärvi	97.9	−2.03	5.02	4.59	0.12	0.96	0.96
Saimaa	96.8	−1.07	3.67	3.52	−0.04	1.11	1.11
Lappajärvi	97.5	0.19	2.87	2.87	0.46	1.51	1.44
Tuusulanjärvi	95.9	0.83	2.92	2.80	0.85	1.41	1.13

for the entire open water period. The impact of the merged observations for chosen lakes was slightly larger (*ca.* 1.5%) than for the SYKE data (compare Tables 1 and 4). In this study, a reliable time series of error statistics could not be obtained due to the small number of lakes. For individual lakes, the mainly positive bias in summer was compensated by the mainly negative bias in spring, so the overall annual bias was quite small. In the EKF-M experiment compared with FR, the absolute value of the bias decreased for Lake Inarijärvi and Lake Saimaa (and the bias even changed sign) and slightly increased for Lake Lappajärvi and Lake Tuusulanjärvi. The RMSE and ESTD are much more demonstrative scores in this type of validation, because errors do not compensate each other. For all lakes, the EKF assimilation procedure resulted in a noticeable decrease in RSME and ESTD. For example, for Lake Inarijärvi ESTD decreased from 4.59°C to 0.96°C. For Lake Tuusulanjärvi, despite the increase in bias and unrealistic temperature profiles in summer, the ESTD for LWST also noticeably decreased (from 2.80°C to 1.13°C). Note that in the EKF-S experiments, the unrealistic behaviour of $\overline{T}$ and T_b did not spoil the LWST analysis.

From cross-validation, the EKF-M experiment was closer to reality than FR. These results are very promising for NWP. The role of early spring observations and errors in $\overline{T}$ and T_b should be studied further in the future. These errors may be relevant for the next winter period. Also, although the simulated deep lake water properties do not influence an atmospheric model, their accurate estimation could be a useful by-product of NWP models for environmental applications (e.g. hydrological or limnological).

5. Conclusions

An algorithm to assimilate LWST observations into the lake model FLake has been developed using the EKF technique. This algorithm propagates information to the water temperature profile in the lake model space. The algorithm considers different lake model regimes: the ice period, the mixed regime and the stratified regime. To avoid the divergence of EKF, the prediction of the B-matrix is

reinitialised when switching between model regimes. The algorithm was applied for two types of data: (1) instrumental measurements of LWST by SYKE and (2) the merged data set, compiled from SYKE measurements, the ice break-up dates observed by SYKE and manually picked MODIS LWST data. With the second type of data, the role of early spring observations was studied. Tests with the SYKE data were performed for 27 lakes in Finland and tests with the merged data for 4 lakes. Experiments comprised a 1 yr period (3 November 2010 to 10 November 2011) and were run offline with forcings from HIRLAM operational forecasts.

In addition to the qualitative evaluation of the results, the impact was calculated for the numerical experiments with SYKE data. Cross-validation was performed for the merged data set. In terms of LWST, for all lakes and all types of data, the EKF experiments are much closer to observations than the free model run, the impact is more than 90%. For all lakes, cross-validation shows a small overall annual bias (less than 0.9°C) for the EKF run. The ESTD with the EKF decreases significantly in comparison with the free model (1–3°C). The experiments with the merged data set demonstrated the important role of the early spring observations to improve the performance of the model in terms of the mean water temperature and the bottom temperature (in the experiments using only the SYKE data they were overestimated). For further evaluation of this behaviour, longer experiments are needed. Numerical experiments revealed a possible model bias in the prognostic equation for the mixed layer depth in the case of the wind mixing in FLake. This problem should be studied in the future. A preliminary study of the behaviour of different elements of the assimilation system was done. Components of Jacobians, the model and background error matrices, the Kalman gain vector showed the annual cycle or time oscillations, no instability was noticed. However, more tests and comparisons with deep water temperature observations are needed. Thus, the task is difficult and the results presented are quite early. Additional numerical experiments using different types of data to study potential problems in the assimilation system (the frequency of

observations, the quality control of remote-sensing data) are planned.

The EKF algorithm used to assimilate LWST data into the lake model FLake showed promising results and was recognised as effective. For the operational implementation however, other parts of the lake DA system (OI analysis, the data quality control, etc.) should also be developed or improved. New sources of data are very important and welcome, in particular space-borne data such as MODIS, Advanced Along-Track Scanning Radiometer (AATSR) and Medium Resolution Imaging Spectrometer (MERIS). The EKF algorithm will be implemented into the externalised surface modelling platform SURFEX (Le Moigne, 2009; Masson et al., 2012) which is widely used in NWP, in climate applications and for monitoring and research purposes. Implementation of SURFEX with FLake and EKF into the operational NWP system HARMONIE/ AROME (Seity et al., 2010) is also planned. The EKF algorithm has strong potential to initialise the lake parameterisation scheme in NWP models and to correct model errors by using LWST observations.

6. Acknowledgements

7. Appendix

Limits of the components of the tangent linear model operator **M**, the linearised observation operator **H**, the background error covariance matrix **B** and the Kalman gain vector **K** for different mixing regimes

Component	Value	Remarks
		Mixed regime
		Matrix **M**
$\frac{\partial \overline{T}'}{\partial \overline{T}^0}$	0.8–0.9 K/K	Almost constant value, less for shallow lakes
		Vector **H**
$\frac{\partial \overline{T}}{\partial \overline{T}}$	1.0 K/K	Constant value
		Matrix **B**
$\text{var}(e_{\overline{T}})$	4.5–4.8 K^2	Almost constant value
		Vector **K** (weights)
for $\overline{T}$	0.82–0.83	Almost constant value
		Stratified regime
		Matrix **M**
$\frac{\partial \overline{T}'}{\partial \overline{T}^0}$	0.5–1.0 K/K	Larger values for deep lakes, smaller for shallow lakes; for shallow lakes the annual cycle has a minimum in midsummer
$\frac{\partial \overline{T}'}{\partial \eta^0}$	(-1.2)–(-0.1) K	Larger absolute values for shallow lakes, smaller for deep lakes; the annual cycle has a maximum in midsummer
$\frac{\partial \overline{T}'}{\partial T_b^0}$	0.01–0.40 K/K	Larger values for shallow lakes, smaller for deep lakes; the annual cycle has a maximum in midsummer
$\frac{\partial \overline{T}'}{\partial C_T^0}$	(-4.5)–(-0.1) K	Larger absolute values for shallow lakes, smaller for deep lakes; the annual cycle has a minimum in midsummer
$\frac{\partial \eta'}{\partial \overline{T}^0}$	(-0.2)–0.2 K^{-1}	Almost zero values in midsummer
$\frac{\partial \eta'}{\partial \eta^0}$	0.1–1.4	Larger values for deep lakes, smaller for shallow lakes; the annual cycle has a maximum in autumn
$\frac{\partial \eta'}{\partial T_b^0}$	(-0.3)–0.5 K^{-1}	Larger absolute values for shallow lakes, smaller for deep lakes; the annual cycle has a minimum in autumn
$\frac{\partial \eta'}{\partial C_T^0}$	(-1.0)–1.0	Almost zero values in midsummer
$\frac{\partial T_b'}{\partial \overline{T}^0}$	(-0.6)–0.6 K/K	The annual cycle with a minimum in spring
$\frac{\partial T_b'}{\partial \eta^0}$	(-2.0)–0.4 K	Close to zero in midsummer
$\frac{\partial T_b'}{\partial T_b^0}$	0.2–1.2 K/K	For shallow lakes an almost constant value of 1.0 K/K, for deep lakes small values in midsummer

Appendix (*Continued*)

Component	Value	Remarks
$\frac{\partial T_b^t}{\partial C_T^0}$	(-6.0)–3.0 K	For shallow lakes low values in midsummer, for deep lakes high values in spring
$\frac{\partial C_T}{\partial \overline{T}^0}$	(-0.03)–0.60 K^{-1}	Almost zero values in midsummer
$\frac{\partial C_T}{\partial \eta^0}$	(-0.01)–0.80	Almost zero values in midsummer
$\frac{\partial C_T}{\partial T_b^0}$	(-1.4)–0.03 K^{-1}	Almost zero values in midsummer
$\frac{\partial C_T}{\partial C_T^0}$	(-0.1)–1.2	For shallow lakes almost zero values in midsummer, for deep lakes an almost constant value of 1.0
Vector **H**		
$\frac{\partial T_{ML}}{\partial \overline{T}}$	1.0–2.6 K/K	The annual cycle has a maximum in midsummer
$\frac{\partial T_{ML}}{\partial \eta}$	0.0–20.0 K	The annual cycle has a maximum in midsummer
$\frac{\partial T_{ML}}{\partial T_b}$	(-1.6)–(-0.1) K/K	The annual cycle has a minimum in midsummer
$\frac{\partial T_{ML}}{\partial C_T}$	0.0–30.0 K	The annual cycle has a maximum in midsummer, the amplitude is stronger for deep lakes
Matrix **B**		
$\mathrm{var}(e_{\overline{T}})$	4.0–16.0 K^2	The annual cycle has a maximum in midsummer, the amplitude is stronger for shallow lakes
$\mathrm{cov}(e_{\overline{T}}, e_{\eta})$	(-2.0)–0.2 K	The annual cycle has a minimum in late summer, the amplitude is stronger for shallow lakes
$\mathrm{cov}(e_{\overline{T}}, e_{T_b})$	(-2.0)–16.0 K^2	The annual cycle has a maximum in late summer, the amplitude is stronger for shallow lakes
$\mathrm{cov}(e_{\overline{T}}, e_{C_T})$	(-0.80)–(-0.01) K	No annual cycle, oscillations
$\mathrm{var}(e_{\eta})$	0.01–0.90	Almost constant small values, but outbreaks in autumn
$\mathrm{cov}(e_{\eta}, e_{T_b})$	(-5.0)–1.0 K	The annual cycle has a minimum in late summer, the amplitude is stronger for shallow lakes
$\mathrm{cov}(e_{\eta}, e_{C_T})$	(-0.05)–0.15	Almost zero values in midsummer
$\mathrm{var}(e_{T_b})$	1.0–45.0 K^2	The annual cycle has a maximum in late summer
$\mathrm{cov}(e_{T_b}, e_{C_T})$	(-2.5)–2.0 K	No annual cycle, oscillations
$\mathrm{var}(e_{C_T})$	0.01–0.18	No annual cycle, oscillations
Vector **K** (weighs)		
for $\overline{T}$	0.20–0.83	The annual cycle has a minimum in midsummer
for η	(-0.06)–0.60 K^{-1}	No annual cycle, oscillations
for T_b	(-0.6)–0.4	No annual cycle, oscillations
for C_T	(-0.05)–0.02 K^{-1}	The annual cycle has a maximum in midsummer

References

Arst, H., Erm, A., Herlevi, A., Kutser, T., Leppäranta, M. and co-authors. 2008. Optical properties of boreal lake waters in Finland and Estonia. *Boreal Environ. Res.* **13**, 133–158.

Balsamo, G., Bouyssel, F. and Noilhan, J. 2004. A simplified bi-dimensional variational analysis of soil moisture from screen-level observations in a mesoscale numerical weather prediction model. *J. Q. Roy. Meteorol. Soc.* **130**, 895–915.

Balsamo, G., Mahfouf, J.-F., Bélair, S. and Deblonde, G. 2007. A land assimilation system for soil moisture and temperature: an information content study. *J. Hydrometeorol.* **8**, 1225–1242.

Balsamo, G., Salgado, R., Dutra, E., Boussetta, S., Stockdale, T. and co-authors. 2012. On the contribution of lakes in predicting near-surface temperature in a global weather forecasting model. *Tellus A.* **64**, 15829. DOI: 10.3402/tellusa.v64i0.15829.

Batrak, Y. 2012. *Lake Climatology Data for Use in Atmospheric Modeling.* Diploma Thesis, Monography, Saint-Petersburg, Russia, 78 (in Russian).

Brasnett, B. 1999. A global analysis of snow depth for numerical weather prediction. *J. Appl. Meteorol.* **38**, 726–740.

Choulga, M., Kourzeneva, E., Zakharova, E. and Doganovsky, A. 2014. Estimation of mean depth of boreal lakes for use in numerical weather prediction and climate modeling. *Tellus A.* **66**, 21295. http://dx.doi.org/10.3402/tellusa.v66.21295.

de Rosnay, P., Drusch, M., Vasiljevic, D., Balsamo, G., Albergel, C. and co-authors. 2013. A simplified extended Kalman filter for the global operational soil moisture analysis at ECMWF. *Q. J. Roy. Meteorol. Soc.* **139**, 1199–1213. DOI: 10.1002/qj.2023.

Donlon, C. J., Martin, M., Stark, J., Robet-Jones, J., Fiedler, E. and co-authors. 2012. The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Rem. Sens. Environ.* **116**, 140–158.

Donlon, C. J., Minnet, P. J., Gentemann, C., Nightingale, T. J., Barton, I. J. and co-authors. 2002. Towards improved validation of satellite sea surface skin temperature measurements for climate research. *J. Clim.* **15**, 353–369.

Draper, C. S., Mahfouf, J.-F. and Walker J. P. 2009. An EKF assimilation of AMSR-E soil moisture into the ISBA land surface scheme. *J. Geophys. Res.* **114**, D20104. DOI: 10.1029/2008JD011650.

Duguay, C. R., Flato, G. M., Jeffries, M. O., Ménard, P., Morris, K. and co-authors. 2003. Ice-cover variability on shallow lakes at high latitudes: model simulations and observations. *Hydrol. Process.* **17**, 3465–3483. DOI: 10.2002/hyp.1394.

Eerola, K., Rontu, L., Kourzeneva, E. and Shcherbak, E. 2010. A study on effects of lake temperature and ice cover in HIRLAM. *Boreal Environ. Res.* **15**, 130–142.

Fiedler, E. K., Martin, M. and Roberts-Jones, J. 2013. An operational analysis of lake water surface temperature. *Tellus A.* in this issue.

Gandin, L. 1965. *Objective Analysis of Meteorological Fields.* Gidrometeoizdat, Leningrad. Translated from Russian, Jerusalem, Israel Program for Scientific Translations, 242 pp.

Golosov, S., Terzhevik, A., Zverev, I., Kirillin, G. and Engelhardt, C. 2012. Climate change impact on thermal oxygen regime of shallow lakes. *Tellus A.* **64**, 17264. DOI: 10.3402/tellusa.v64i0.17264.

Gustafsson, N. 1985. Development of mesoscale analysis scheme for nowcasting and very short range forecasting. In: *Proceedings of the ECMWF Workshop on High Resolution Analysis*, Reading, 24–26 June 1985, pp. 183–212. (Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK.).

Haines, K. 2010. Ocean data assimilation. In: *Data Assimilation: Making Sense of Observations* (eds. W. A. Lahoz, B. Khattatov, and R. Ménard). Springer, Berlin, pp. 517–547. DOI: 10.1007/978-3-540-74703-1_20.

Hess, R. 2001. Assimilation of screen-level observations by variational soil moisture analysis. *Meteorol. Atmos. Phys.* **77**, 145–154.

Homleid, M. 2009. Sea surface temperature and sea ice concentration in HIRLAM and HARMONIE. *HIRLAM Newsletter N55 part B*, pp. 55–64. Online at: http://www.hirlam.org

Houser, P. R., De Lannoy, G. J. M. and Walker, J. P. 2010. Land surface data assimilation. In: *Data Assimilation: Making Sense of Observations* (eds. W. A. Lahoz, B. Khattatov, and R. Ménard). Springer, Berlin, pp. 549–597. DOI: 10.1007/978-3-540-74703-1_21.

Jonas, T., Terzhevik, A. Y., Mironov, D. V. and Wüest, A. 2003. Radiatively driven convection in an ice-covered lake investigated by using temperature microstructure technique. *J. Geophys. Res.* **108**(C6), 3183. DOI: 10.1029/2002JC001316.

Kalnay, E. 2003. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, Cambridge, UK, 341 pp.

Kheyrollah Pour, H., Duguay, C. R., Solberg, R. and Rudjord, Ø. 2014a. Impact of satellite-based lake surface observations on the initial state of HIRLAM – part I: remotely-sensed lake water temperature observations. *Tellus A.* **66**, 21534. Online at: http://dx.doi.org/10.3402/tellusa.v66.

Kheyrollah Pour, H., Rontu, L., Duguay, C., Eerola, K. and Kourzeneva, E. 2014b. Impact of satellite-based lake surface observations on the initial state of HIRLAM – part II: analysis of lake surface temperature and ice cover. *Tellus A.* **66**, 21395. Online at: http://dx.doi.org/10.3402/tellusa.v66.

Kheyrollah Pour, H., Duguay, C. R., Martynov, A. and Brown, L. C. 2012. Simulation of surface temperature and ice cover of large northern lakes with 1-D models: a comparison with MODIS satellite data and *in-situ* measurements. *Tellus A.* **64**, 17614. DOI: 10.3402/tellusa.v64i0.17614.

Kirillin, G. 2010. Modeling the impact of global warming on water temperature and seasonal mixing regimes in small temperate lakes. *Boreal Environ. Res.* **15**, 279–293.

Kourzeneva, E., Asensio, H., Martin, E. and Faroux, S. 2012a. Global gridded dataset of lake coverage and lake depth for use in numerical weather prediction and climate modeling. *Tellus A.* **64**, 15640. DOI: 10.3402/tellusa.v64i0.15640.

Kourzeneva, E., Martin, E., Batrak, Y. and Le Moigne, P. 2012b. Climate data for parameterisation of lakes in numerical weather prediction models. *Tellus A.* **64**, 17226. DOI: 10.3402/tellusa.v64i0.17226.

Krinner, G. 2003. Impact of lakes and wetlands on boreal climate. *J. Geophys. Res.* **108**(D16), 4520. DOI: 10.1029/2002JD002597.

Le Moigne, P., Boone, A., Calvet, J.-C., Decharme, B., Faroux, S. and co-authors. 2009. *SURFEX Scientific Documentation*. CNRM, Météo France, France, 221 pp. Online at: http://www.cnrm.meteo.fr/surfex/

Lewis, W. M., Jr. 1983. A revised classification of lakes based on mixing. *Can. J. Fish. Aquat. Sci.* **40**, 1779–1787.

MacCallum, S. N. and Merchant, C. J. 2012. Surface water temperature observations of large lakes by Optimal Estimation. *Can. J. Rem. Sens.* **38**(1), 25–45. DOI: 10.5589/m12-010.

Mahfouf, J.-F. 1991. Analysis of soil moisture from near-surface parameters: a feasibility study. *J. Appl. Meteorol.* **30**, 1534–1547.

Mahfouf, J.-F., Bergaoui, K., Draper, C., Boussel, F., Taillefer, F. and co-authors. 2009. A comparison of two off-line soil analysis schemes for assimilation of screen level observations. *J. Geophys. Res.* **114**, D08105. DOI: 10.1029/2008JD011077.

Martynov, A., Sushama, L., Laprise, R., Winger K. and Dugas, B. 2012. Interactive lakes in the Canadian regional climate model, version 5: the role of lakes in the regional climate of North America. *Tellus A.* **64**, 16226. DOI: 10.3402/tellusa.v64i0.16226.

Masson, V., Le Moigne, P., Martin, E., Faroux, S., Allas, A. and co-authors. 2012. The SURFEX7.2 land and ocean surface platform for coupled offline simulation of earth surface variables and fluxes. *Geosci. Model Dev.* **5**, 3771–3851. DOI: 10.5194/gmdd-5-3771-2012.

Mironov, D. 2008. *Parameterization of Lakes in Numerical Weather Prediction. Description of a Lake Model*. COSMO Technical Report 11, Deutscher Wetterdienst, Offenbach am Main, Germany.

Mironov, D., Ritter, B., Schultz, J.-P., Buchhold, M., Lange, M. and co-authors. 2012. Parameterisation of sea and lake ice in numerical weather prediction models of the German Weather Service. *Tellus A.* **64**, 17330. DOI: 10.3402/tellusa.v64i0.17330.

Navascues, B. 1997. *Analysis of 2 m Temperature and Relative Humidity*. HIRLAM Technical Report No 28, SMHI, Norrköping, Sweden.

Potes, M., Costa, M. J., da Silva, J. C. B., Silva, A. M. and Morais, M. 2011. Remote sensing of water quality parameters over Alqueva Reservoir in the South of Portugal. *Int. J. Rem. Sens.* **32**(12), 3373–3388. DOI: 10.1080/01431161003747513.

Rhodin, A., Kucharski, F., Callies, U., Eppel, D. P. and Wergen, W. 1999. Variational soil moisture analysis from screen-level atmospheric parameters: application to a short-range weather forecast model. *Q. J. Roy. Meteorol. Soc.* **125**, 2427–2448.

Rontu, L., Eerola, K., Kourzeneva, E. and Vehviläinen, B. 2012. Data assimilation and parametrisation of lakes in HIRLAM. *Tellus A.* **64**, 17611. DOI: 10.3402/tellusa.v64i0.17611.

Salgado, R. and Le Moigne, P. 2010. Coupling of the FLake model to the surfex externalized surface model. *Boreal Environ. Res.* **15**, 231–244.

Samuelsson, P., Kourzeneva, E. and Mironov, D. 2010. The impact of lakes on the European climate as simulated by a regional climate model. *Boreal Environ. Res.* **15**, 113–129.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P. and co-authors. 2010. The AROME-France convective-scale opera-tional model. *Mon. Weather. Rev.* **139**, 976–990. DOI: 10.1175/2010MWR3425.1.

Stroud, J., Lesht, B. M., Schwab, D. J., Beletsky, D. and Stein, M. L. 2009. Assimilation of satellite images into a sediment transport model of Lake Michigan. *Water Resour. Res.* **45**, W02419. DOI: 10.1029/2007WR006747.

Undén, P., Rontu, L., Järvinen, H., Lynch, P., Calvo, J. and co-authors. 2002. The HIRLAM-5 scientific documentation. Online at: http://hirlam.org

Zhang, Z., Beletsky, D., Schwab, D. J. and Stein, M. L. 2007. Assimilation of current measurements into a circulation model of Lake Michigan. *Water Resour. Res.* **43**, W11407.