RMetS
Royal Meteorological Society

Evaluation of NWP results for wintertime nocturnal boundary-layer temperatures over Europe and Finland

Evgeny Atlaskin[a,b,c*] and Timo Vihma[a]
[a]*Finnish Meteorological Institute, Helsinki, Finland*
[b]*University of Helsinki, Finland*
[c]*Russian State Hydrometeorological University, St Petersburg, Russia*
*Correspondence to: E. Atlaskin, Atmospheric Research, Finnish Meteorological Institute, Erik Palmenin Aukio 1, Helsinki 00101, Finland. E-mail: evgeny.atlaskin@gmail.com

Four operational numerical weather prediction (NWP) models were evaluated in winter conditions against (a) synoptic observations in Europe, (b) observations at a 48 m high micrometeorological mast in Sodankylä, northern Finland, and (c) observations at the Helsinki Testbed stations: (i) to evaluate the skills of the models to compute nocturnal 2 m air temperature (T2m) and the temperature inversion; and (ii) to distinguish between the T2m bias and the subgrid-scale spatial variability of T2m. The models were (1) the Integrated Forecast System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF), (2) the High Resolution Limited Area Model (HIRLAM), (3) the Applications of Research to Operations at Mesoscale (AROME) developed by Météo-France, and (4) the Global Forecasting System (GFS) of the US National Center for Environmental Predictions (NCEP). The results demonstrated a T2m bias increasing with decreasing temperature and strengthening temperature inversion. When a strong temperature inversion was observed in Sodankylä, the models underestimated it, whereas in near-neutral conditions the stratification was overestimated. Comparison of observed and modelled 3 h temperature tendencies showed that the T2m tendency in the models was on average only 17–20% of the observed one. The warm bias in T2m forecast in Sodankylä during periods of observed temperature inversion partly resulted from a warm bias in the initial conditions. This was due to problems in data assimilation in IFS and HIRLAM, in initialization in AROME, and in either or both procedures in GFS. In particular, the IFS data assimilation increased the T2m bias. Evaluation of modelled T2m against grid-averaged T2m observed at Helsinki Testbed demonstrated that the T2m model error dominated over the spatial variability of observed T2m. This suggests that over an almost flat terrain horizontal resolution is not a major factor for the accuracy of T2m forecast at low T2m typically associated with temperature inversions. **Copyright © 2012 Royal Meteorological Society**

Key Words: numerical weather prediction; two-metre temperature; stable boundary layer; subgrid-scale variability; mesoscale observational network; model verification

Received 2 May 2011; Revised 15 November 2011; Accepted 14 December 2011; Published online in Wiley Online Library

Citation: Atlaskin E, Vihma T. 2012. Evaluation of NWP results for wintertime nocturnal boundary-layer temperatures over Europe and Finland. *Q. J. R. Meteorol. Soc.* DOI:10.1002/qj.1885

1. Introduction

Deterministic forecast of 2 m air temperature (T2m) over snow-covered land in conditions of nocturnal stable boundary layer (SBL) has often been associated with a large temperature bias in numerical weather prediction (NWP) models (e.g. Järvenoja, 2005; Maas *et al.*, 2008; Tastula and Vihma, 2011). The largest errors typically appear in the forecast of T2m in wintertime when the temperature is lowest and stratification is strongest.

The main processes responsible for the formation of SBL are (1) cooling of the surface due to a negative radiation budget (Sun *et al.*, 2003), (2) warm-air advection over a cold surface (Vihma *et al.*, 2003), and (3) subsidence (Yi *et al.*, 2001). A variety of subgrid-scale processes may occur in a SBL, e.g. intermittent turbulence, gravity waves, low-level jets and density currents (Mahrt *et al.*, 1998; Mahrt, 1999), which in NWP models are only partly resolved or entirely parametrized. To perform computationally affordable simulations, the description of boundary layer physics is simplified in NWP models. Calculation of turbulent fluxes near the surface is usually based on the Monin–Obukhov similarity theory, derived assuming a horizontal homogeneity and vertically constant turbulent fluxes. Under strong static stability, however, the constant-flux layer may not exist or may not reach the height of the lowest model level (Mahrt, 1999). In this case the Monin–Obukhov similarity theory is violated. Further, full integration of radiative transfer equations is time consuming and typically not affordable in operational runs. Instead, radiative fluxes are computed by either simplified fast schemes (e.g. Savijärvi, 1990) or by full integration schemes with coarse horizontal and/or time resolutions (e.g. Morcrette *et al.*, 2008b). These simplifications impose inaccuracies in calculation of the long-wave radiation and turbulent fluxes.

Although the problem of a large T2m bias is well known to modellers, it is still important to quantify the performance of state-of-the-art NWP models in nocturnal winter conditions. Although some NWP models provide T2m values separately for different terrain/vegetation types within a grid cell, in most models T2m represents an average over the grid cell, whereas a local observation is a point value, which complicates their comparison (e.g. Hanna and Yang, 2001). In SBL, due to weak wind and advection, the influence of local terrain properties on the surface energy balance and further on the local T2m amplifies, thus increasing the horizontal temperature variation. Consequently, observed T2m becomes representative only for a limited area surrounding the station and having the same surface properties. Hence, even if the forecast for a grid-averaged T2m is perfect, its comparison against a local observation shows an apparent error that often increases with strengthening stability. To diminish such an apparent error, the comparison should be made against observations averaged within the model grid cell. Mesoscale observational networks allowing it are, however, rare.

Several studies have been devoted to the evaluation of the performance of mesoscale models (Cox *et al.*, 1998; Hanna and Yang, 2001; Zhong and Fast, 2003; Steeneveld *et al.*, 2008). All the studies showed that the models underestimate diurnal temperature cycle amplitude and near-surface temperature stratification at night.

In this paper, results of four operational NWP models run at European and US centres are evaluated in wintertime nocturnal conditions exploiting observations from the whole of Europe and Finland, where low near-surface temperatures and temperature inversion are often observed. The models are (1) the Integrated Forecast System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF), (2) the High Resolution Limited Area Model (HIRLAM), (3) the Applications of Research to Operations at Mesoscale (AROME) developed by Météo-France and (4) the Global Forecasting System (GFS) of the US National Center for Environmental Predictions (NCEP).

Our objectives are:

- to evaluate the skills of operational NWP models to predict nocturnal T2m and the temperature inversion in winter;
- to evaluate the errors in data assimilation, forecast initialization, and temporal evolution of the forecasts;
- to distinguish between the contributions of inaccuracies in modelling of SBL and the limited representativeness of observations to the errors in T2m.

The observations employed to evaluate model results are described in section 2, and the four models selected are presented in section 3. Evaluation results for T2m forecasts are presented in section 4. Section 5 focuses on the effect of the temperature inversion on the accuracy of T2m forecasts and analyses. Comparisons of temporal evolution of observed and simulated near-surface temperatures are also presented in section 5. In section 6 a mesoscale network of observations is used to distinguish between the model errors and subgrid-scale variability of T2m. Discussion and conclusions are presented in section 7.

2. Observations

The evaluation was done for the winter season from 1 December 2009 to 1 March 2010 as it was the latest one by the time the study was started. Weather conditions in Scandinavia and Eastern Europe during the winter period were predominantly governed by high-pressure systems, weak near-surface winds and low near-surface temperatures, which are typically associated with temperature inversion and absence of low-level clouds. In December 2009, snow or ice covered most Scandinavian land areas and inland waters. In January and February 2010, sea ice covered the Gulf of Finland and Gulf of Bothnia, and central and eastern parts of Europe were partly covered by snow. Over most of Finland the winter was coldest since 1986–1987, with temperatures 1–5°C lower than the mean of the 30-year normal period of 1971–2000.

The following three datasets were exploited in the evaluation of the model results: (1) synoptic (SYNOP) observations from Europe were utilized to evaluate 24 h forecasts for T2m; (2) observations from Sodankylä, northern Finland, were exploited to study the relationship of T2m errors and the strength of temperature inversion, to evaluate the modelled inversion, and to estimate T2m bias in the first guess, analysis and initialized forecast; and (3) observations at Helsinki Testbed stations were used in comparing the modelled and observed grid-averaged T2m, to estimate the role of subgrid-scale variability and the horizontal grid resolution on the accuracy of T2m forecast.

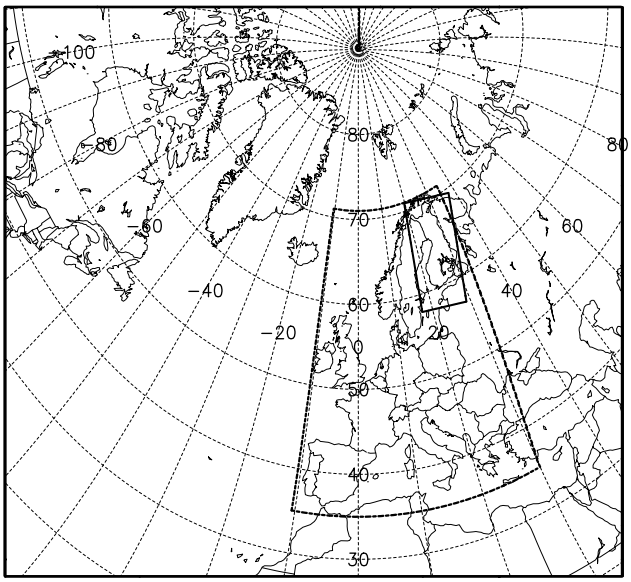

Figure 1. A domain embracing all EWGLAM stations (area with dotted boundaries), and the forecast domains of FMI HIRLAM (large rectangular) and FMI AROME (small rectangular).

2.1. *Synoptic observations*

The evaluation of T2m forecasts was done using data from the stations included in the European Working Group on Limited Area Modelling (EWGLAM) list. EWGLAM stations are located in the region of 10.5°W to 30°E, 35°N to $72.5°N (Figure 1), and are commonly used to validate NWP models. The evaluation of AROME results was done for the stations located within its smaller integration domain.

2.2. *Sodankylä observations*

Observations were made at the Arctic Research Centre of the Finnish Meteorological Institute (FMI-ARC), which is located in Sodankylä, northern Finland at 67.36°N, $26.63°E, 179 m above sea level. A 48 m high micrometeorological mast is deployed on a sandy soil in a Scots pine forest, having a moderate density of trees 10–12 m tall. Air temperature was measured at different heights by Pentronic PT100 sensor. Only temperatures measured at the heights of 3 m (T3m) (closest to the height of 2 m) and 32 m (T32m) were used in this study.

The terrain around FMI-ARC is almost flat and mostly covered with mixed spruce and deciduous forest. The 200 m wide River Kittinen flows 220 m west of the mast. In wintertime the river is covered with ice and snow. There is a peat bog of 2.5 km^2 located nearly 250 m west of the mast. A more detailed description of the site can be found in Batchvarova *et al.* (2001) and Atlaskin and Kangas (2006). The instruments in Sodankylä and Helsinki Testbed stations were regularly calibrated by FMI staff. In addition, all the data exploited in the study passed the quality control, which included admissible range checks, as well as removal of spikes and non-changing values.

2.3. *Helsinki Testbed*

Helsinki Testbed consists of synoptic and road weather stations, masts and sounding stations (Koskinen *et al.*, 2011). In the region (Figure 2), the surface elevation

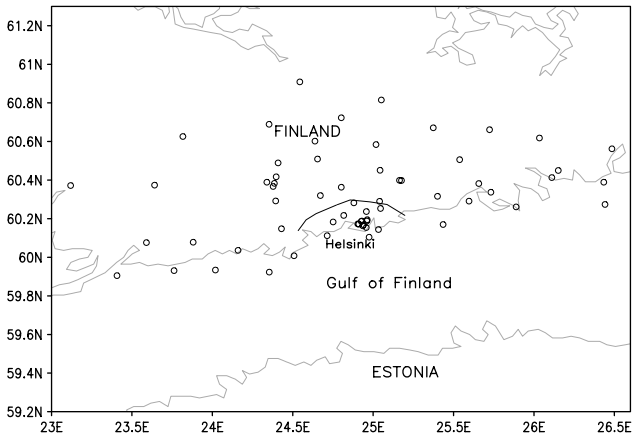

Figure 2. Locations of Helsinki Testbed stations (dots) within and outside motorway Ring III (black solid line).

gradually increases to the northnorthwest with a gradient of approximately 2 m per 1000 m. The surface near the southern coast of Finland consists of mixed deciduous and coniferous forest, lakes, agricultural fields, small rocky islands, as well as Helsinki and smaller towns and villages. The urban environment of Helsinki and neighbouring towns is mostly concentrated south of the Ring III motorway (Figure 1). The minimum distance between the stations is approximately 2 km. In the present study, data from 36 weather stations were utilized. T2m was observed using Vaisala WXT510 sensors.

3. Models

Operational NWP models have a variety of numerical and physics parametrization schemes to account for soil, snow, vegetation, orography and various atmospheric properties, each one imposing a degree of uncertainty in T2m forecasts. Here we summarize the physical parametrizations of the four models applied. Among them only AROME is a non-hydrostatic model. All four models employ a terrain following hybrid sigma-pressure vertical coordinate. Basic numerical properties of the models are presented in Table 1, and the integration domains of the limited-area models HIRLAM and AROME run operationally at the Finnish Meteorological Institute are displayed in Figure 1. HIRLAM acquires boundary conditions from IFS, whereas AROME acquires both initial and boundary conditions from HIRLAM. The resolution of IFS was changed on 26 January 2010 (Table 1). Also corrections in short-wave radiation interaction with clouds and corrections in handling of land surface parameters were implemented (http://www.ecmwf.int/products/changes/horizontal_resolution_2009/). However, we did not detect any effect on near-surface temperatures. Hence statistical calculations were made for a single dataset that contains results from both IFS versions.

3.1. *Parametrization schemes*

The physical parametrizations essential for SBL are summarized in Table 2. HIRLAM and AROME apply an ABL scheme based on the turbulent kinetic energy (TKE), whereas IFS and GFS apply first-order closure schemes. In IFS, HIRLAM and AROME the vertical mixing in the ABL

Table 1. Basic properties of the NWP models applied in the study.

Model	Domain, integration scheme	Horizontal grid spacing	Horizontal resolution of validated fields	Number of levels in vertical	Height of the lowest model level (m)
IFS cycle 35r3 updated to 36r1 26 Jan. 2010	Global, spectral	25 km reduced 26 Jan. 2010 to 16 km	0.25°	91	10
HIRLAM 7.1	LAM, finite-difference	15 km	0.15°	60	30
AROME 33h1	LAM, spectral	2.5 km	0.022°	40	30
GFS	Global, spectral	56 km	0.5°	64	20

Table 2. Physical parametrization schemes of the NWP models applied in the study.

Model	Mixing in PBL	Mixing in the surface layer	Radiation	Soil heat flux
IFS[a]	First-order closure scheme (Louis *et al.*, 1982; Nieuwstadt, 1984)	M-O similarity profile, stability functions of Dyer (1974) and Högström (1988)	RRTM for SW and LW fluxes	4-layer scheme based on diffusion equation
HIRLAM[b]	TKE-l scheme (Lenderink and Holtslag, 2004)	M-O similarity profile relationships, stability function of Louis (1979)	Savijärvi (1990) fast radiation scheme	2-layer force-restore scheme (Noilhan and Planton, 1989)
AROME[c]	TKE-l scheme with Bougeault and Lacarrère (1989) mixing length	M-O similarity profile relationships, stability function of Louis (1979)	RRTM for LW and Morcrette (1991) scheme for SW radiation flux	2-layer force-restore scheme (Noilhan and Planton, 1989)
GFS[d,e]	Non-local first-order scheme (Troen and Mahrt, 1986; Hong and Pan, 1996)	M-O similarity profile relationships (Miyakoda and Sirutis, 1986) with modified stability functions for very stable and very unstable cases	Chou and Suarez (1999) scheme for SW and RRTM (Mlawer *et al.*, 1997) for LW radiation fluxes	4-layer scheme based on diffusion equation (Koren *et al.*, 1999; Ek *et al.*, 2003)

[a] http://ecmwf.int/research/ifsdocs/
[b] http://hirlam.org/index.php?option=com_docman&task=doc_download&gid=270&Itemid=70
[c] http://www.cnrm.meteo.fr/arome/doc/arodoc.pdf
[d] http://www.meted.ucar.edu/nwp/pcu2/index.htm
[e] http://www.emc.ncep.noaa.gov/officenotes/newernotes/on442.pdf

is simulated using a combination of eddy diffusivity and mass flux schemes (EDMF). Orographic effects, such as wave trapping, blocking and wave propagation, are taken into account in IFS, HIRLAM and GFS, whereas AROME does not have a separate parametrization for orographic effects. In all the models, the sea-surface temperature (SST) is prescribed on the basis of the surface analysis and kept constant during the forecast. A single-layer snow scheme is used for computation of the snow amount and thermodynamics in all models, although with differences in the detailed parametrizations. Below the snow, the number of layers in the soil scheme varies between two and four (Table 2). All the models apply a few land tiles (six in GFS and three in the other models) with separate surface temperatures. In addition, HIRLAM includes an ice tile and AROME a town tile. The radiation schemes applied in the four models vary from a simple, fast scheme (Savijärvi, 1990) applied in HIRLAM to the sophisticated RRTM scheme with the Monte Carlo Independent Column Approximation approach (Morcrette *et al.*, 2008a) for cloud–radiation interactions, applied in IFS. The condensation schemes in IFS, HIRLAM and GFS are based on Sundqvist (1978), whereas AROME employs a subgrid scheme for warm-phase and ice-phase clouds. The precipitable water content is computed as a diagnostic variable in IFS, GFS and HIRLAM, and as a prognostic variable in AROME. Convection schemes applied in the models are 1D schemes, based on Tiedtke (1989) in IFS, Kain and Fritsch (1990) in HIRLAM and Pan and Wu (1994) in GFS. Only a shallow convection scheme is applied in AROME.

3.2. *Model output*

As a low T2m and a temperature inversion are basically observed at night time, the evaluation for EWGLAM and Helsinki Test bed stations was done using the forecasts valid at 0000 UTC.

Estimation of the models' performance with respect to data assimilation, initialization and forecast for Sodankylä mast was done also for daytime as the solar radiation at 67°N is very limited in winter. IFS analysis files were available every 6 h, whereas forecast files were available only at 0000 and 1200 UTC. Therefore, T2m bias in the analysis was calculated with 6 h intervals, T2m bias in the first guess was calculated for 0600 and 1800 UTC and T2m bias in initialized forecast was calculated for 0000 and 1200 UTC. In the other models, T2m biases were calculated for 0000 and 1200 UTC. AROME does not have its own data assimilation system and performs initialization using only HIRLAM analysis. From GFS, only first-guess fields and initialized forecasts were available.

The forecast length was selected to be 24 h, which was the maximum in AROME operational cycle and covers the diurnal cycle. Hence the accuracy of T2m forecast does not only depend on the model performance at observed night-time temperature inversions.

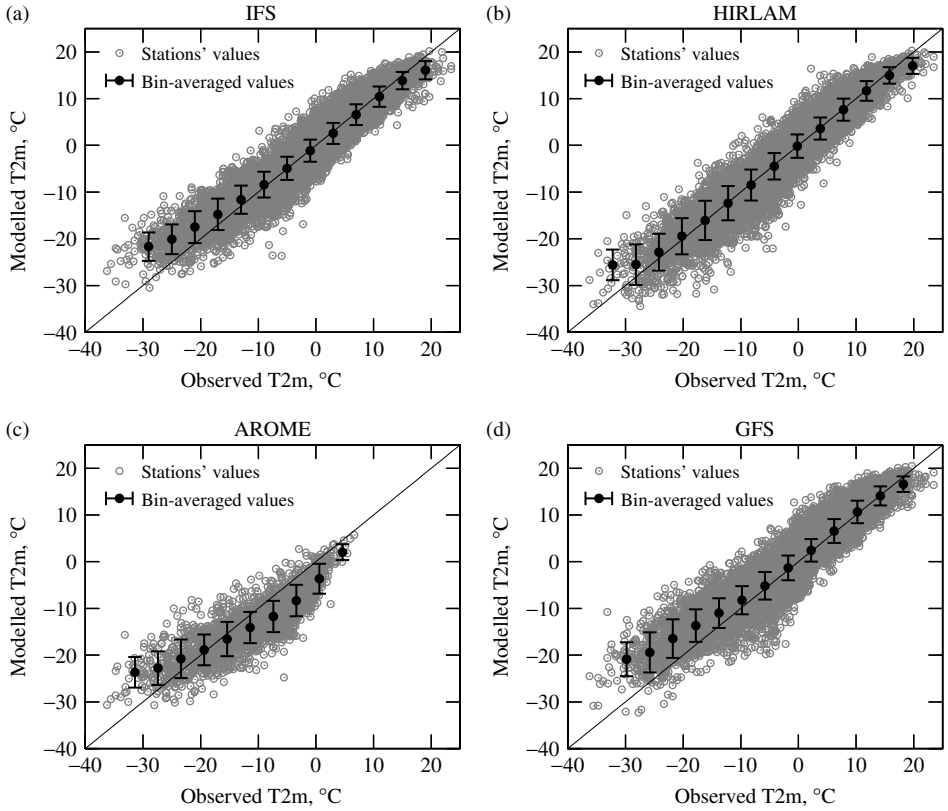

Figure 3. Scatterplots of observed (EWGLAM stations) and modelled T2m. The black dots denote the simulated T2m averaged over 4°C intervals of observed temperatures, and the vertical bars indicate the standard deviations. The grey circles denote values at the stations.

In all the models, T2m was diagnosed from the surface temperature (Ts) and the lowest model-level temperature. However, we did not find a notable difference between Ts and T2m in the model simulations compared to T2m model−measurement bias. This implies that the errors in T2m diagnostics did not contribute much to the T2m bias.

4. Evaluation against synoptic observations

4.1. Methods

The model results were bilinearly interpolated to the EWGLAM station locations. Height correction to the modelled temperature was done using dry adiabatic lapse rate. Moreover, to minimize the uncertainty associated with the effects of orography, the data from stations located higher than 300 m above sea level were not used. IFS, GFS and HIRLAM results for T2m were verified against observations done at all EWGLAM station, whereas AROME results were verified against the observations done at the EWGLAM stations located within its domain. To demonstrate the statistical dependence of T2m errors on the observed temperature, the data were divided into three classes: low (T2m < -20°C), moderate (-20°C $\leq$ T2m < 10°C) and high temperatures (T2m ≥ 10°C).

4.2. Results

Comparisons of observed and modelled T2m at EWGLAM stations revealed that all the models overestimated T2m at low observed T2m; the lower the observed T2m, the stronger was the overestimation (Figure 3). Inversely, IFS, HIRLAM and GFS predominantly underestimated T2m when the observed value exceeded 10°C; the higher the observed T2m, the stronger was the underestimation. The corresponding T2m bias and root mean squared error (RMSE) are given in Table 3.

AROME strongly underestimated T2m within the range of observed T2m from -15 to 0°C. The cases of strongest underestimation by GFS occurred when T2m ranged from -10 to 0°C. The distribution of points in Figure 3 roughly corresponds to the latitudinal distribution, with low temperatures and positive biases mainly observed in northern latitudes and high temperatures and negative biases mainly observed in low latitudes.

Figure 4 displays the time series of T2m averaged over all EWGLAM stations as well as T2m averaged over the stations located within the AROME domain. The temperature averaging over the AROME domain was done to estimate the performance of all models for this northern region, where low near-surface temperatures and temperature inversions are often observed. Averaged over all EWGLAM stations, IFS and HIRLAM well reproduced the observed T2m, whereas GFS overestimated it. The results for the AROME domain revealed that positive T2m biases usually corresponded to the observed T2m minima. In general, both HIRLAM and AROME systematically and significantly underestimated T2m within AROME domain, which was related to strong underestimation of Ts.

In all the models, the largest positive T2m bias was associated with a strong decrease of observed T2m (for observed T2m tendencies less than -0.8 K h^{-1}, the mean T2m bias among four models varying from 3.6°C in HIRLAM to 5.7°C in AROME), whereas the largest negative T2m bias was associated with a strong increase

Table 3. Mean T2m bias and root mean squared error (RMSE) at EWGLAM stations for the classes of low, moderate and high temperatures.

Model	T2m < −20°C		−20 ≤ T2m ≤ 10°C		T2m > 10°C	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
IFS	4.5	5.8	0.0	2.4	−0.6	2.0
HIRLAM	1.3	4.4	−0.1	2.6	−0.1	1.8
AROME	3.1	5.2	−2.0	4.8	–	–
GFS	6.0	7.4	0.8	2.8	0.3	2.1

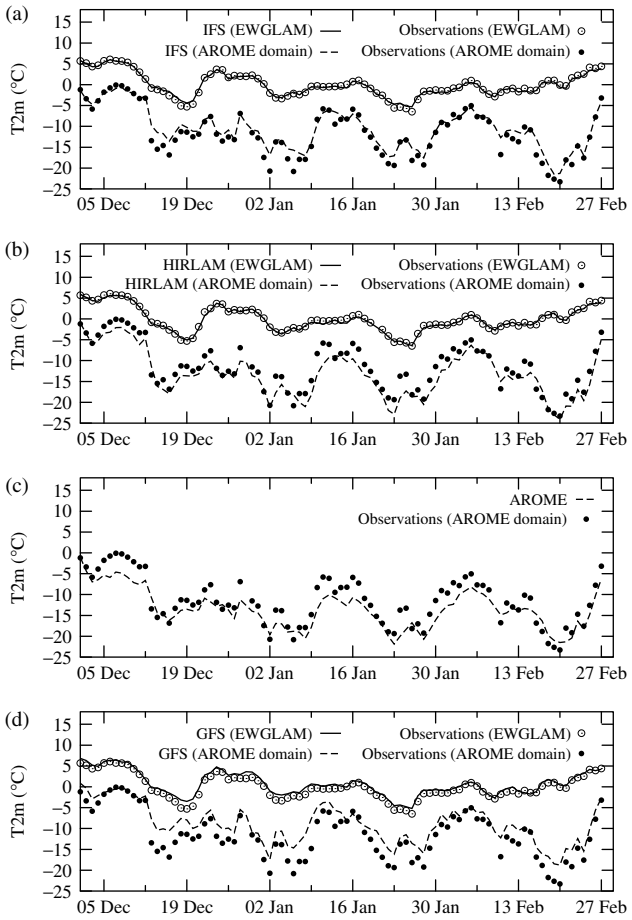

Figure 4. Time series of observed (open circles) and modelled (solid lines) T2m averaged over all EWGLAM stations and of observed (dots) and modelled (dashed lines) T2m averaged over stations located within the AROME domain.

of observed T2m (for observed T2m tendencies above 0.8 K h^{-1}, the mean T2m bias varying among the models from −6.7°C in AROME to −2°C in IFS). T2m tendency was calculated for the interval from 1200 to 0000 UTC, which covers the period of decreasing solar radiation at all EWGLAM stations. In HIRLAM and AROME the strongest underestimations corresponded to the observed T2m maxima. GFS predominantly overestimated T2m both at negative and positive tendencies.

The results presented above demonstrate that the models underestimated the range of variability of T2m. In HIRLAM, a small positive T2m bias associated with low observed temperatures (Figure 3(b)) resulted from the fact that the model strongly and systematically underestimated T2m in the northern regions. Negative T2m bias in the northern regions reversed into a positive bias predominantly under a rapid decrease of the observed T2m. In the AROME domain, the temporal variation of T2m in AROME resembles the variation in HIRLAM (Figure 3(b, c)), which was partly due to the fact that AROME received its initial and boundary conditions from HIRLAM.

5. Evaluation against Sodankylä mast measurements

To investigate the relationship of a low T2m, an associated positive T2m bias and a temperature inversion, observations from the inversion layer are needed. Hence Sodankylä data were applied to complement the evaluation based on EWGLAM data.

5.1. *Methods*

The temperature gradient was calculated on the basis of observations of T3m and T32m. The modelled gradients were calculated for practically the same layer, using diagnostic T2m and an upper-level temperature. For HIRLAM and AROME, the latter was the temperature at the lowest model level, located at a height of 31 m. To obtain the gradient from the same layer, in IFS the temperature at the second-lowest model level at a height of 31 m was applied, whereas for GFS we interpolated the values at two adjacent levels to the level of 32 m.

To evaluate the dependence of T2m bias on the observed temperature gradient, the latter was divided into three classes: negative ($dT/dz < 0°C\ m^{-1}$), moderately positive ($0°C\ m^{-1} \leq dT/dz < 0.1°C\ m^{-1}$) and strongly positive ($dT/dz > 0.1°C\ m^{-1}$).

5.2. *Results*

5.2.1. *24 h forecasts*

In the canopy layer the temperature profile was practically neutral, resulting in a minor difference between T3m and T2m. It is worth noting that vertical temperature gradient calculated from the difference between T32 and T3 m is an overestimation for the canopy layer, where neutral stratification prevailed, and an underestimation for the layer above, where a stratified temperature profile was often observed at night. The T2m bias was calculated as the difference between the predicted T2m and the measured T3m.

The comparison of the simulated T2m and the T3m observed at Sodankylä yielded results qualitatively similar to those obtained for EWGLAM stations: all models systematically overestimated T2m at a low observed T3m – the lower the observed temperature, the stronger was the overestimation. The T2m bias increased with increasing vertical temperature gradient (VTG) based on the 48 m high mast measurements at heights of 3 and

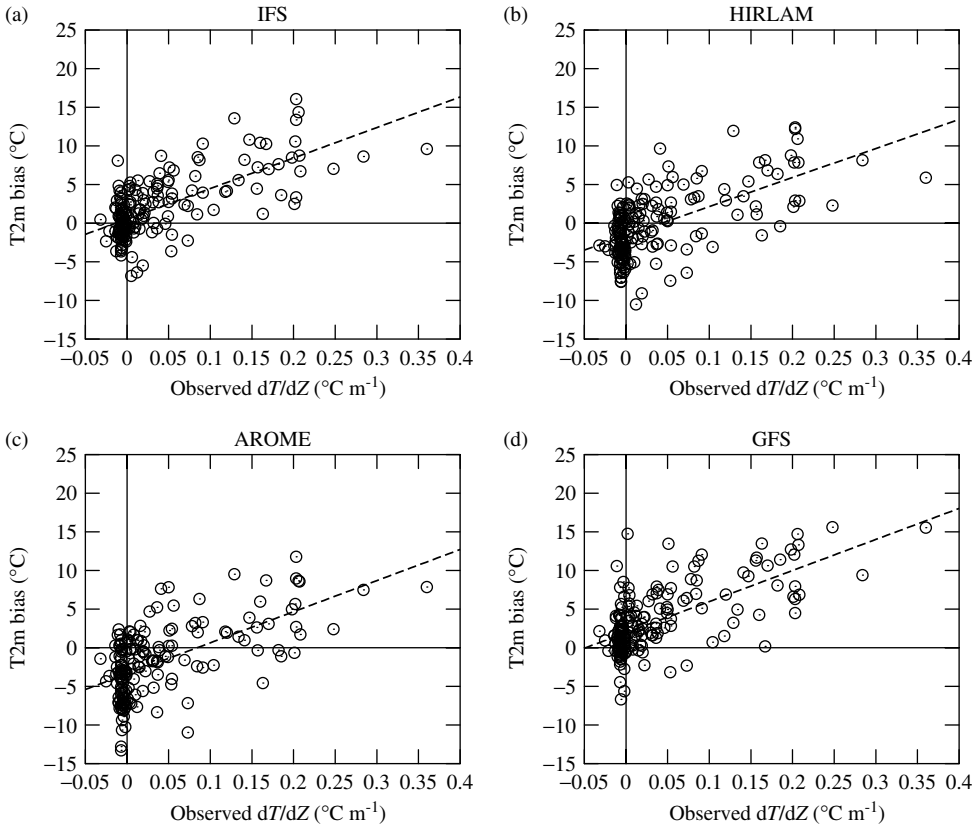

Figure 5. T2m bias relative to the observed temperature gradient in Sodankylä.

Table 4. Mean T2m bias and temperature gradient bias as percentage of observed temperature gradient in Sodankylä for three classes of vertical temperature gradient.

Model	$dT/dz < 0$ K m^{-1}		$0 \leq dT/dz \leq 0.1$ K m^{-1}		$dT/dz > 0.1$ K m^{-1}	
	T2m bias (° C)	dT/dz bias (%)	T2m bias (° C)	dT/dz bias (%)	T2m bias (° C)	dT/dz bias (%)
IFS	−0.1	129	2.3	116	7.8	−93
HIRLAM	−2.5	99	0.4	−104	5.2	−91
AROME	−4.5	1381	−0.9	479	3.8	−47
GFS	1.1	69	4.2	−95	8.6	−82

32 m (Figure 5). Cases with strong temperature inversion were typically associated with low T3m values. Multiple regression analysis performed for T2m bias as a function of both observed T3m and temperature gradient demonstrated relatively strong correlation that equalled 0.78 for AROME, 0.72 for IFS and GFS and 0.69 for HIRLAM.

All the models failed to reproduce the observed VTG (Figure 6). T2m bias and VTG bias divided by the observed VTG, calculated for different VTG categories, are given in Table 4. All the models to a variable extent overestimated VTG when the observed value was negative or slightly stable, and underestimated VTG in conditions of large observed values. In conditions of an observed positive VTG (Figure 6(a)), IFS was the only model that simulated a negative temperature stratification, with the temperature gradient down to -0.05°C m^{-1}. Under an observed strongly positive VTG, AROME results for VTG agreed best with the observations; however, the model simulated mainly moderately positive VTG and significantly overestimated VTG in near-neutral temperature stratification. Both HIRLAM and GFS simulated basically near-neutral temperature stratification.

The evaluation results based on EWGLAM stations suggested that the night-time T2m bias was related to an underestimation of temporal variations of T2m. The comparison was, however, only based on 24 h forecasts. Comparison of observed and modelled 3-hourly T2m tendencies during 24 h long simulations demonstrated that the models strongly underestimated the T2m tendency. In IFS and GFS it was on average only 17% and in HIRLAM and AROME 20% of the observed tendency. The models, however, simulated better the temperature tendency at the height of 32 m, which in IFS and HIRLAM practically equalled the observed one, and in AROME and GFS was 50% of the observed one. This is, however, associated with the decrease of the observed temporal variations with height.

5.2.2. Data assimilation and initialization

It turned out that problems also exist in data assimilation and model initialisation. Table 5 addresses cases of observed strongly positive temperature gradient (>0.1 K m^{-1}) and presents the T2m bias in (a) 6 h forecasts that are used as the first guess for data assimilation, (b) the models

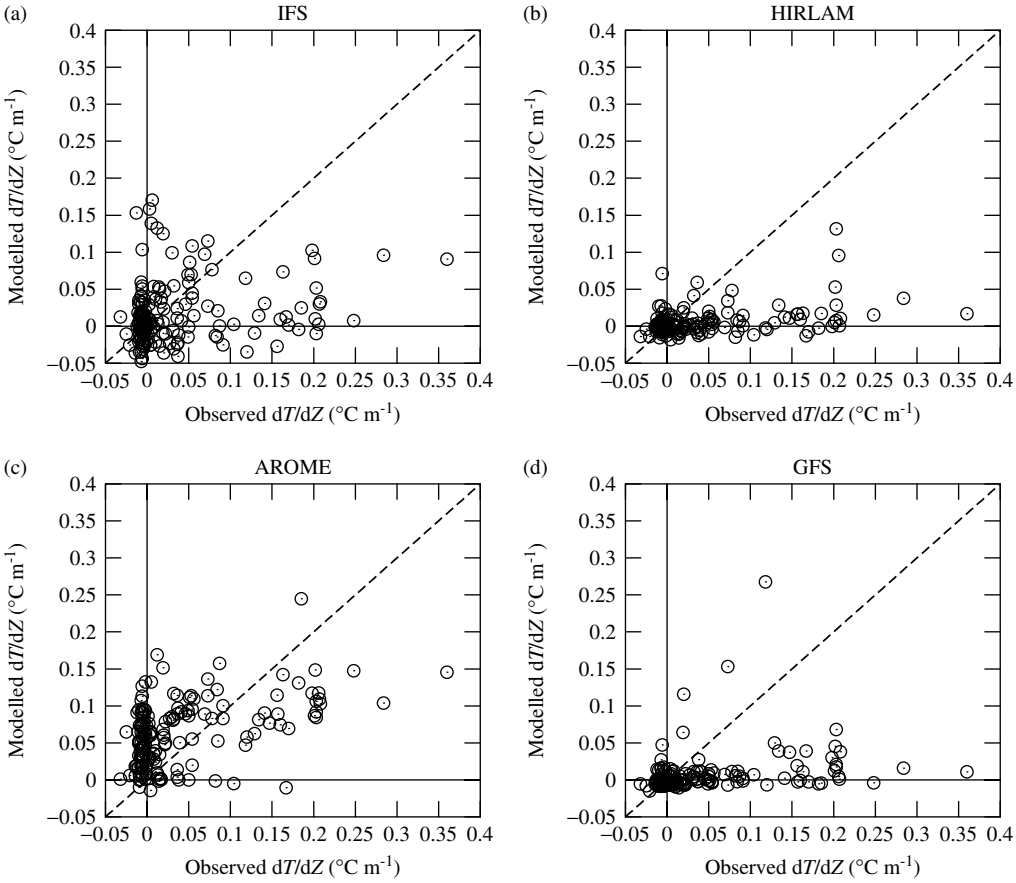

Figure 6. Modelled temperature gradient relative to the observed temperature gradient in Sodankylä.

Table 5. Mean T2m bias in 6 h forecast (first guess), analysis and initialized forecast, calculated for cases of observed strongly positive temperature gradient ($dT/dz > 0.1$ K m^{-1}).

Model	First-guess T2m bias	Analysis T2m bias	Initialized forecast T2m bias	24 h forecast T2m bias
IFS	5	6.8	5.2	7.8
HIRLAM	4.2	3	2.9	5.2
AROME	–	3	3.1	3.8
GFS	6.7	–	7.3	8.6

analyses, (c) the initialized forecast and (d) 24 h forecast. In IFS, AROME and GFS, errors in the initial values were on average as large as in the 24 h forecasts, and in HIRLAM the bias in the initial value was approximately half of the bias in the 24 h forecast. In IFS, the T2m bias was largest in the analysis, whereas the errors in the first guess and the initialized forecast were practically equal to each other. This means that the IFS data assimilation increases the T2m bias, but it is again decreased by the model initialization. In HIRLAM analysis and initialized forecast the T2m bias was almost the same and less than the bias in the first guess, suggesting that the data assimilation improves the situation. AROME uses HIRLAM analysis and initializes with a slightly larger T2m bias. In GFS, the large T2m bias of initialized forecast is 0.6°C larger than the bias of the first guess, suggesting a harmful effect of either the analysis or initialization or both.

The above results demonstrate that in conditions of observed temperature inversions the forecast errors partly result from the errors in the initial fields. Further, the low density of synoptic observations does not allow capturing local variability of T2m, which results in uncertainties both in the initial model values and in the results of model–measurement comparison. This will be addressed in the next section.

6. Evaluation against Helsinki Testbed observations

6.1. *Methods*

T2m model error was calculated as follows:

$$\Delta\mathrm{T2m} = \mathrm{T2m_{fc}} - \frac{1}{N}\sum_{i=1}^{N}\mathrm{T2m}^{i}_{\mathrm{obs}} \qquad (1)$$

where N is the number of observation stations located within the grid cell. The subscripts 'fc' and 'obs' refer to forecast and observations, respectively. The following criteria were applied in selecting the grid cells for averaging of T2m: (i) the minimum number of stations within the grid cell is three for GFS, having the largest grid size, and two for the rest of the models; (ii) there are at least two stations within the grid cell with a mutual distance at least half of the mean grid cell length (otherwise the subgrid-scale variability is not necessarily well detected by the stations); (iii) the fraction of stations located on the open sea or lakes approximately equals the fraction of water surface in the grid cell. The ΔT2m provides the best available estimate for the true model bias associated with errors in various model schemes. The more stations there are in a grid cell, the more reliable

Table 6. Mean T2m bias (°C), T2m model error (ΔT2m, °C) and relative T2m error (f_{mod}, °C/°C) calculated for the classes of low, moderate and high temperatures observed at Helsinki Testbed stations.

Model		T2m < −15°C	−15°C ≤ T2m ≤ 0°C	T2m > 0°C
IFS	T2m bias	4.1	0.9	−0.4
	ΔT2m	3.2	0.8	−0.6
	f_{mod}	4.7	1	−2.5
HIRLAM	T2m bias	−0.6	−2	−0.4
	ΔT2m	−0.1	−1.7	−0.2
	f_{mod}	−0.5	−4.3	−1
AROME	T2m bias	−0.6	−2.7	−1.1
	ΔT2m	−1.6	−3.4	−1.6
	f_{mod}	−4.4	−6.4	−2.1
GFS	T2m bias	4.7	1.8	−0.2
	ΔT2m	3.5	1.4	−0.1
	f_{mod}	2.3	0.9	−0.2

is the evaluation of the model's performance. The number of grid cells satisfying the above criteria was, however, very limited, reaching five for GFS, four for IFS, three for HIRLAM and two for AROME. In any case, the dataset allows defining statistical correlation between ΔT2m and the grid-averaged observed T2m. In addition, the modelled T2m was interpolated to the station locations to calculate T2m bias averaged over the Helsinki Testbed area, with all the station data used.

Spatial variability of observed T2m within the model's grid cell represents an uncertainty in the calculation of T2m bias that can be positive for one station and negative for another one in the same grid cell. To compare the T2m model error with the observed subgrid-scale variability, the relative T2m error was calculated as follows:

$$f_{mod} = \frac{\Delta T2m}{|\Delta T2m_{obs}|_{max}} \quad (2)$$

where $|\Delta T2m_{obs}|_{max}$ is the maximum absolute difference of observed T2m within the grid cell. If $|f_{mod}| > 1$, the model error is the main source for the T2m bias at individual stations.

Averaging observed T2m in the grid cell reduces the population of values less than −20°C. In addition, temperatures above 10°C were not observed during the winter at the network's stations. Therefore, the observed temperature averaged over the grid cell was divided into the following temperature classes: low (T2m < −15°C), moderate (−15°C ≤ T2m ≤ 0°C) and high (T2m > 0°C).

6.2. *Results*

The results indicated that a low observed T2m corresponded to the largest positive T2m bias and the largest positive T2m model error (ΔT2m) in IFS and GFS, whereas HIRLAM and AROME predominantly underestimated T2m both over land and sea ice, resulting in a negative T2m bias and T2m model error (Table 6). The results for ΔT2m were qualitatively similar to the results for T2m bias, calculated for the network stations, and in accordance with the results obtained for EWGLAM stations and Sodankylä.

Figure 7 illustrates the distribution of relative T2m error, based on Eq. (2), averaged over 4°C intervals as well as a quadratic fit of the error. The latter is calculated to extrapolate the results for f_{mod} to temperatures bellow −20°C, which were rarely observed at the Testbed stations. The extrapolation is supported by the results of the model−measurement comparison obtained for EWGLAM and Sodankylä stations. The T2m model error generally dominates over the spatial variability of observed T2m within the model grid cells. For IFS and GFS the relative T2m error is larger at lower observed temperatures, implying that the domination of ΔT2m over the observed subgrid-scale spatial variability strengthens with decreasing observed temperature. A large negative T2m bias in HIRLAM and AROME resulted in large negative values of the relative T2m error at moderate temperatures (Table 6). However, the quadratic fit, together with the EWGLAM and Sodankylä evaluation results, suggests that the T2m bias may turn positive at observed temperature bellow −20°C.

The distribution of subgrid-scale variability of observed T2m is presented in Figure 8. Accordingly, the values are based solely on observations, but calculated for the different grids of the five models. T2m spatial variability within HIRLAM and AROME grid cells is not sensitive to the observed temperature, implying that the decrease of the magnitude of relative T2m error (Figure 7(b, c)) is associated with a smaller ΔT2m. The spatial variability of observed T2m peaks at moderate temperatures of −10 to −15°C. Such variability is associated with a large difference between temperatures observed on land and sea stations during the period when sea ice was absent or thin. In such conditions, the spatial variability is largest when T2m is lowest. The variability of observed T2m within IFS and GFS grid cells decreases at low temperatures, which were typically observed when more compact and thicker sea ice was present. HIRLAM and AROME grid cells only included land stations, which decreased the subgrid-scale T2m variability.

7. Discussion and conclusions

Compared to many previous evaluations of NWP results, our approach included the following important aspects: (1) the use of state-of-the-art global and limited-area NWP models for European and Nordic regions; (2) a 3-month-long evaluation period, which included many cases of stable stratification and allowed us to determine statistical dependencies of errors on the observed conditions; and (3) minimization of the uncertainty associated with subgrid-scale variability of the observed T2m. At all observation sites (EWGLAM, Helsinki Testbed stations and Sodankylä) our results demonstrated a T2m bias increasing with decreasing temperature and strengthening temperature inversion. The vertical profiles of observed and modelled temperatures in the lowermost 32 m layer differed significantly. When a strong temperature inversion was observed in Sodankylä the models underestimated it, whereas in observed slightly unstable conditions all models, particularly AROME, produced predominantly stable stratification.

Comparison of modelled T2m with grid-averaged T2m observed at Helsinki Testbed stations yielded results similar to those obtained for Sodankylä and EWGLAM stations. The T2m model error (ΔT2m) was positive and systematically increased with decreasing observed T2m. ΔT2m dominated over the spatial variability of observed T2m, suggesting that the model horizontal resolution is not a major factor for the accuracy of T2m forecast over an almost flat terrain in

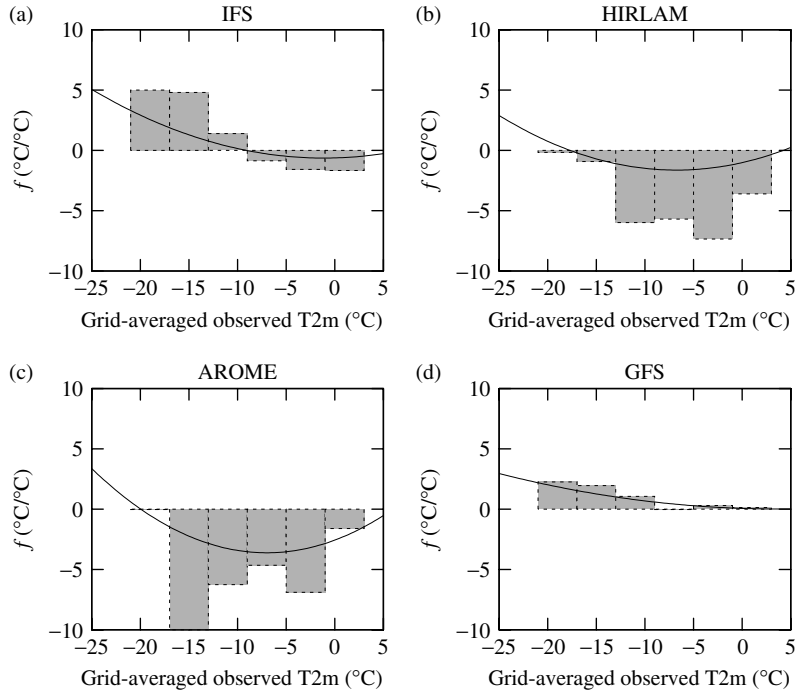

Figure 7. Distribution of relative T2m error averaged over 4°C intervals (top edges of the bars) and approximated with a quadratic function (solid curve) in the range of grid-averaged observed T2m for IFS (a), HIRLAM (b), AROME (c) and GFS (d) for the Helsinki Testbed stations.

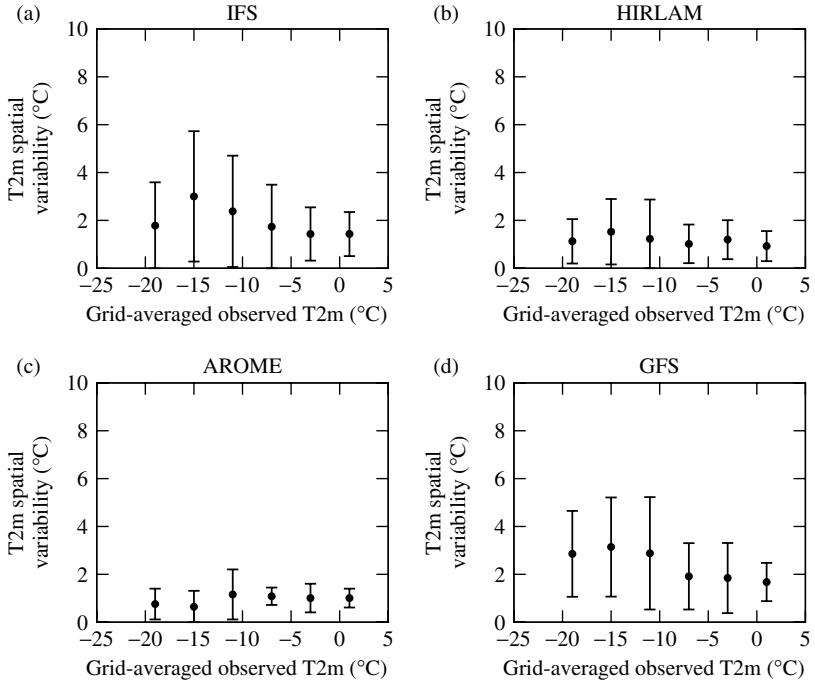

Figure 8. Distribution of subgrid-scale spatial variability of observed T2m averaged over 4°C intervals for the Helsinki Testbed stations.

cold winter conditions typically associated with temperature inversions (although the situation may be different if there is both open sea and snow/ice-covered areas in the grid cell; Vihma, 1995). Analogous conclusions were obtained in the Antarctic SBL simulations by Tastula and Vihma (2011).

Positive T2m bias and underestimation of temperature inversion were related to strong underestimation of temporal variation of T2m in the models; on average, the modelled T2m tendency did not exceed 20% of the observed tendency. During observed temperature inversions, the models strongly overestimated the initial T2m, revealing a serious problem in data assimilation in IFS and a smaller problem in HIRLAM (some bias remains after the data assimilation), and a major problem either in data assimilation or forecast initialization or both in GFS. In AROME, the initialization increased the bias of the analysis performed by HIRLAM. The problem of data assimilation in IFS is in line with Lüpkes *et al.* (2010), who discovered large warm and moist biases in the ABL over the Arctic Ocean in the ERA Interim reanalysis of the ECMWF, although the comparisons were made against observations utilized in the data assimilation.

Factors leading to large biases in T2m, the temperature gradient and the temporal temperature variation were not concretely studied here, but our results are basically in accordance with previous studies. Although the four models applied in our study had large differences in the physical parametrization schemes (Table 2), the main results were qualitatively similar in the sense that under lowest temperatures a positive bias strongly dominated. A possible explanation is that, to avoid numerical instability potentially resulting from thermal decoupling of the surface and atmosphere, operational models tend to overestimate the level of background turbulence in very stable conditions. This is a common problem for climate models as well (Tjernström *et al.*, 2005), and was also revealed in the Global Energy and Water Cycle Experiment (GEWEX) Atmospheric Boundary Layer Study (GABLS) experiments (Cuxart *et al.*, 2006; Steeneveld *et al.*, 2006; Svensson *et al.*, 2011).

T2m forecasts could probably be improved by introducing more vertical levels (Hanna and Yang, 2001), a separate parametrization scheme for a vegetation layer (Steeneveld *et al.*, 2008) and more detailed treatment of snow and ice thermodynamics (Cheng and Vihma, 2002). Further, many NWP models still apply relatively simple radiation schemes, but in conditions of stable stratification the turbulent fluxes are small and the relative importance of radiative fluxes on T2m accordingly increases, suggesting the use of more sophisticated schemes, such as Morcrette *et al.* (2008b) applied in IFS. The main challenges remain, however, in the ABL scheme. Our results stress the importance of a proper treatment of decoupling conditions. Approaches suggested for that include the Quasi-Normal-Scale Elimination method (Sukoriansky *et al.*, 2005), a parametrization of minimum eddy diffusivity as a function of subgrid-scale orography (Savijärvi, 2009), and the introduction of a new stability parameter also taking into account the effects of gravity waves and Earth rotation (Zilitinkevich and Esau, 2005). More model experiments are, however, needed to better understand the practical benefits of these new approaches.

Acknowledgements

We thank Sami Niemelä and Markku Kangas for providing us with observational data, and Ekaterina Kourzeneva and Laura Rontu for fruitful discussions. The study was supported by the EU FP6 'ModObs' project, a St Petersburg grant for young researchers and by the Academy of Finland, grant 131723.

References

Atlaskin E, Kangas M. 2006. Sodankylä data utilization for HIRLAM verification and 1D model studies. *HIRLAM Newsl.* **51**: 103–112.

Batchvarova E, Gryning S-E, Hasager CB. 2001. Regional fluxes of momentum and sensible heat over a sub-arctic landscape during late winter. *Bound.-Lay. Meteorolol.* **99**: 489–507.

Bougeault P, Lacarrère P. 1989. Parameterization of orography-induced turbulence in a meso-beta scale model. *Mon. Weather Rev.* **117**: 1872–1890.

Cheng B, Vihma T. 2002. Idealized study of a 2-D coupled sea-ice/atmosphere model during warm-air advection. *J. Glaciol.* **48**: 425–438.

Chou M-D, Suarez MJ. 1999. A shortwave radiation parameterization for atmospheric studies. *NASA Tech. Memo.* **15**(104606): pp. 51. http://gmao.gsfc.nasa.gov/pubs/docs/Chou136.pdf

Cox R, Bauer BL, Smith T. 1998. A mesoscale model intercomparison. *Bull. Am. Meteorol. Soc.* **79**: 265–283.

Cuxart J, Holtslag AAM, Beare RJ, Bazile E, Beljaars A, Cheng A, Conangla L, Ek M, Freedman F, Hamdi R, Kerstein A, Kitagawa H, Lenderink G, Lewellen D, Mailhot J, Mauritsen T, Perov V, Schayes G, Steeneveld G-J, Svensson G, Taylor P, Weng W, Wunsch S, Xu K-M. 2006. Single-column model intercomparison for a stably stratified atmospheric boundary layer. *Bound.-Lay. Meteorol.* **118**: 273–303.

Dyer AJ. 1974. A review of flux–profile relationships. *Bound.-Lay. Meteorol.* **7**: 363–372.

Ek MB, Mitchell KE, Lin Y, Rogers E, Grunmann P, Koren V, Gayno G, Tarpley JD. 2003. Implementation of Noah land-surface model advances in the NCEP operational mesoscale Eta model. *J. Geophys. Res.* **108**: 8851, DOI: 10.1029/2002JD003296.

Hanna SR, Yang R. 2001. Evaluation of mesoscale models' simulations of near-surface winds, temperature gradients, and mixing depths. *J. Appl. Meteorol.* **40**: 1095–1104.

Hogström U. 1988. Non-dimensional wind and temperature profiles in the atmospheric surface layer: A re-evaluation. *Bound.-Lay. Meteorol.* **42**: 55–78.

Hong S-Y, Pan H-L. 1996. Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.* **124**: 2322–2339.

Järvenoja S. 2005. 'Problems in predicted HIRLAM T2m in winter, spring and summer'. In *Proceedings of the SRNWP/HIRLAM Workshop on Surface Processes, Surface Assimilation and Turbulence*, Norkköping, Sweden, 15–17 September 2004. Available: http://srnwp.met.hu/workshops/Norrkoping_2004/05_SJa.pdf

Kain JS, Fritsch JM. 1990. A one dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.* **47**: 2784–2802.

Koren V, Schaake J, Mitchell K, Duan Q-Y, Chen F, Baker J. 1999. A parameterization of snowpack and frozen ground intended for NCEP weather and climate models. *J. Geophys. Res.* **104**(D16): 19569–19585.

Koskinen JT, Poutiainen J, Schultz DM, Joffre S, Koistinen J, Saltikoff E, Gregow E, Turtiainen H, Dabberdt WF, Damski J, Eresmaa N, Göke S, Hyvärinen O, Järvi L, Karppinen A, Kotro J, Kuitunen T, Kukkonen J, Kulmala M, Moisseev D, Nurmi P, Pohjola H, Pylkkö P, Vesala T, Viisanen Y. 2011. The Helsinki testbed: a mesoscale measurement, research, and service platform. *Bull. Am. Meteorol. Soc.* **92**: 325–343.

Lenderink G, Holtslag AAM. 2004. An updated length-scale formulation for turbulent mixing in clear and cloudy boundary layers. *Q. J. R. Meteorol. Soc.* **130**: 3405–3427.

Louis JF. 1979. A parametric model of vertical eddy fluxes in the atmosphere. *Bound.-Lay. Meteorol.* **17**: 187–202.

Louis JF, Tiedtke M, Geleyn J-F. 1982. 'A short history of the operational PBL parametrization at ECMWF'. In *Proceedings of ECMWF Workshop on Boundary Layer Parametrization*, Reading, November 1981. ECMWF: Reading, UK.

Lüpkes C, Vihma T, Jakobson E, König-Langlo G, Tetzlaff A. 2010. Meteorological observations from ship cruises during summer to the central Arctic: a comparison with reanalysis data. *Geophys. Res. Lett.* **37**: L09810, DOI: 10.1029/2010GL042724.

Maas CF, Baars J, Wedam G, Grimit E, Steed R. 2008. Removal of systematic model bias on a model grid. *Weather Forecast.* **23**: 438–459.

Mahrt L. 1999. Stratified atmospheric boundary layers. *Bound.-Lay. Meteorol.* **90**: 375–396.

Mahrt L, Sun J, Blumen W, Delany T, Oncley S. 1998. Nocturnal boundary-layer regimes. *Bound.-Lay. Meteorol.* **88**: 255–278.

Miyakoda K, Sirutis J. 1986. *Manual of the E-physics*. Geophysical Fluid Dynamics Laboratory, Princeton University: Princeton, NJ.

Mlawer EJ, Taubman SJ, Brown PD, Iacono MJ, CLough SA. 1997. Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.* **102D**: 16663–16682.

Morcrette J-J. 1991. Radiation and cloud radiative properties in the European Centre for Medium Range Weather Forecast Forecasting system. *J. Geophys. Res.* **96D**: 9121–9132.

Morcrette J-J, Mozdzynski G, Leutbecher M. 2008a. Impact of a new radiation package, McRad, in the ECMWF integrated forecasting system. *Mon. Weather Rev.* **136**: 4773–4798.

Morcrette J-J, Mozdzynski G, Leutbecher M. 2008b. A reduced radiation grid for the ECMWF integrated forecasting system. *Mon. Weather Rev.* **136**: 4760–4772.

Nieuwstadt FTM. 1984. The turbulent structure of the stable, nocturnal boundary layer. *J. Atmos. Sci.* **41**: 2202–2216.

Noilhan J, Planton S. 1989. A simple parameterization of land surface processes for meteorological models. *Mon. Weather Rev.* **117**: 536–549.

Pan H-L, Wu W-S. 1994. Implementing a mass-flux convective parameterization package for the NMC Medium Range Forecast

Model. *Preprints, 10th Conf. on Numerical Weather Prediction*, Portland, OR. American Meteorological Society: Boston, MA. 96–98.
Savijärvi H. 1990. Fast radiation parameterization schemes for mesoscale and short-range forecast models. *J. Appl. Meteorol.* **29**: 437–447.
Savijärvi H. 2009. Stable boundary layer: model parameterizations for local and larger scales. *Q. J. R. Meteorol. Soc.* **135**: 914–921.
Steeneveld GJ, van de Wiel BJH, Holtslag AAM. 2006. Modelling the Arctic nocturnal stable boundary layer and its coupling to the surface. *Bound.-Lay. Meteorol.* **118**: 357–378.
Steeneveld GJ, Mauritsen T, de Brujin EIF, de Arellano JV-G, Svensson G, Holtslag AAM. 2008. Evaluation of limited-area models for the representation of the diurnal cycle and contrasting nights in CASES-99. *J. Appl. Meteorol. Climatol.* **47**: 869–887.
Sukoriansky S, Galperin B, Staroselsky I. 2005. A quasinormal scale elimination model of turbulent flows with stable stratification. *Phys. Fluids* **17**: 085107, DOI: 10.1063/1.2009010.
Sun J, Burns SP, Delany AC, Oncley SP, Horst TW, Lenschow DH. 2003. Heat balance in the nocturnal boundary layer during CASES-99. *J. Appl. Meteorol.* **42**: 1649–1666.
Sundqvist H. 1978. A parameterization scheme for non-convective condensation including prediction of cloud water content. *Q. J. R. Meteorol. Soc.* **104**: 677–690.
Svensson G, Holtslag AAM, Kumar V, Mauritsen T, Steeneveld GJ, Angevine WM, Bazile E, Beljaars A, de Bruijn EIF, Cheng A, Conangla L, Cuxart J, Ek M, Falk MJ, Freedman F, Kitagawa H, Larson VE, Lock A, Mailhot J, Masson V, Park S, Pleim J, Söderberg S, Weng W, Zampieri M. 2011. Evaluation of the diurnal cycle in the atmospheric boundary layer over land as represented by a variety of single-column models: the second GABLS experiment. *Bound.-Lay. Meteorol.* **140**: 177–206.
Tastula E-M, Vihma T. 2011. WRF model experiments on the Antarctic atmosphere in winter. *Mon. Weather Rev.* **139**: 1279–1291.
Tiedtke M. 1989. A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Mon. Weather Rev.* **117**: 1779–1800.
Tjernström M, Zagar M, Svensson G, Cassano JJ, Pfeifer S, Rinke A, Wyser K, Dethloff K, Jones C, Semmler T, Shaw M. 2005. Modelling the Arctic boundary layer: an evaluation of six ARCMIP regional-scale models using data from the SHEBA project. *Bound.-Lay. Meteorol.* **117**: 337–381.
Troen IB, Mahrt L. 1986. A simple model of the atmospheric boundary layer: sensitivity to surface evaporation. *Bound.-Lay. Meteorol.* **37**: 129–148.
Vihma T. 1995. Subgrid parameterization of surface heat and momentum fluxes over polar Oceans. *J. Geophys. Res.* **100**: 22625–22646.
Vihma T, Hartmann J, Lüpkes C. 2003. A case study of an on-ice air flow over the Arctic marginal sea ice zone. *Bound.-Lay. Meteorol.* **107**: 189–217.
Yi C, Davis KJ, Berger BW, Bakwin PS. 2001. Long-term observations of the dynamics of the continental planetary boundary layer. *J. Atmos. Sci.* **58**: 1288–1299.
Zhong S, Fast J. 2003. An evaluation of the MM5, RAMS, and meso-eta models at subkilometer resolution using VTMX field campaign data in the Salt Lake valley. *Mon. Weather Rev.* **131**: 1301–1322.
Zilitinkevich SS, Esau IN. 2005. Resistance and heat/mass transfer laws for neutral and stable planetary boundary layers: old theory advanced and re-evaluated. *Q. J. R. Meteorol. Soc.* **131**: 1863–1892.

