Boundary-Layer Meteorology (2004) 117: 337–381
DOI 10.1007/s10546-004-7954-z

MODELLING THE ARCTIC BOUNDARY LAYER: AN EVALUATION OF SIX ARCMIP REGIONAL-SCALE MODELS USING DATA FROM THE SHEBA PROJECT

MICHAEL TJERNSTRÖM[1,*], MARK ŽAGAR[1], GUNILLA SVENSSON[1], JOHN J. CASSANO[2], SUSANNE PFEIFER[5], ANNETTE RINKE[3], KLAUS WYSER[4], KLAUS DETHLOFF[3], COLIN JONES[4], TIDO SEMMLER[5] and MICHAEL SHAW[3]

[1]*Department of Meteorology, Stockholm University, Stockholm, Sweden;* [2]*Cooperative Institute for Research in the Environmental Sciences and Program in Atmospheric and Oceanic Sciences, University of Colorado, BO, U.S.A.;* [3]*Alfred Wegener Institute for Polar and Marine Research, Postdam, Germany;* [4]*Swedish Meteorological and Hydrological Institute, Norrköping, Sweden;* [5]*Max-Planck Institute for Meteorology, Hamburg, Germany*

(Received in final form 16 December 2004)

Abstract. A primary climate change signal in the central Arctic is the melting of sea ice. This is dependent on the interplay between the atmosphere and the sea ice, which is critically dependent on the exchange of momentum, heat and moisture at the surface. In assessing the realism of climate change scenarios it is vital to know the quality by which these exchanges are modelled in climate simulations. Six state-of-the-art regional-climate models are run for one year in the western Arctic, on a common domain that encompasses the Surface Heat Budget of the Arctic Ocean (SHEBA) experiment ice-drift track. Surface variables, surface fluxes and the vertical structure of the lower troposphere are evaluated using data from the SHEBA experiment. All the models are driven by the same lateral boundary conditions, sea-ice fraction and sea and sea-ice surface temperatures. Surface pressure, near-surface air temperature, specific humidity and wind speed agree well with observations, with a falling degree of accuracy in that order. Wind speeds have systematic biases in some models, by as much as a few metres per second. The surface radiation fluxes are also surprisingly accurate, given the complexity of the problem. The turbulent momentum flux is acceptable, on average, in most models, but the turbulent heat fluxes are, however, mostly unreliable. Their correlation with observed fluxes is, in principle, insignificant, and they accumulate over a year to values an order of magnitude larger than observed. Typical instantaneous errors are easily of the same order of magnitude as the observed net atmospheric heat flux. In the light of the sensitivity of the atmosphere–ice interaction to errors in these fluxes, the ice-melt in climate change scenarios must be viewed with considerable caution.

Keywords: Arctic climate, Climate, Climate model, Numerical modelling.

1. Introduction

Model projections of anthropogenic climate change indicate large climate sensitivity in the Arctic (e.g. IPCC, 2001). The ensemble average Arctic

[*] E-mail: michaelt@misu.su.se

warming in 19 CMIP (Coupled Model Intercomparison Project, Meehl et al., 2000) simulations is about 2.5 times larger than the average global warming (Räisänen, 2001). However, the global general circulation models (GCM) have problems in reproducing even the current Arctic climate; they are generally too warm, have systematic biases in surface pressure fields and the surface radiative fluxes vary widely between models (Walsh et al., 2002). Consequently, the inter-model spread in the CMIP climate-warming scenarios is much larger in the Arctic than elsewhere on Earth (Räisänen, 2001). Difficulties in simulating the Arctic climate relate directly to an insufficient understanding of several strong feedback processes. The large climate sensitivity in models is largely due to the strongly positive snow-and-ice/albedo feedback: warming reduces the ice and/or snow cover, thereby reducing the surface albedo, further enhancing the warming. How climate models handle snow and sea ice is therefore critical. Battisti et al. (1997) showed that lack of physical detail in the description of ice processes inhibits realistic representation of the natural variability in the Arctic climate. It also causes significant errors in global weather forecast models (Beesley et al., 2000).

The Arctic environment, with its semi-permanent sea ice, sets up unique atmospheric boundary-layer conditions. The annual cycle is very large, while the diurnal cycle, which influences the boundary-layer structure at many mid-latitude locations, is often absent. During Arctic winter, the snow-covered ice insulates the atmosphere from the relatively warm ocean. Combined with the absence of solar warming, strong longwave surface cooling facilitates the formation of long-lasting surface inversions with strongly stable conditions. The Arctic boundary layer (ABL) is stably stratified about 75% of the time (Persson et al., 2002) and turbulence in very stable conditions is generally poorly understood (Mahrt, 1998). The longevity of the stable conditions makes the interplay between gravity waves and turbulence relatively more important (Zilitinkevich, 2002). During summer the ice melts, which efficiently regulates the low-level atmospheric temperature. Additional energy input melts the snow and ice rather than heating the surface, while energy loss results in the freezing of water rather than the cooling of the surface.

Long periods of stable ABL conditions in winter are interspersed with periods of near-neutral conditions, forced by longwave radiation (Persson et al., 1999, 2002) directly related to boundary-layer clouds; also a known problem for models. In the Arctic winter, liquid water droplets are present in a sizeable fraction even at very low temperatures (Beesly et al., 2000; Intrieri et al., 2002). Unusual vertical structures in summer, with layering and decoupling from the surface, were described in Curry (1986) and Curry et al. (2000). Clouds play an important role for the surface radiative fluxes, determining the net longwave radiation and regulating incoming solar radiation in summer. Over the Arctic pack ice, in contrast to mid-latitude oceans, clouds often lead to surface warming (Intrieri et al., 2002).

Many physical processes in climate models are not resolved and therefore need to be parameterised. Development of parameterisations always involves an empirical component. Detailed process observations in the Arctic are, however, sparse and consequently the ensemble of observations forming the empirical basis for the development of reliable parameterisations may therefore be inadequate. It is important to develop, test and evaluate such schemes using *in situ* measurements from the Arctic. Until quite recently, this was difficult due to the lack of adequate data representing a reasonable ensemble of Arctic conditions. This situation is improving, with new experiments in the Arctic, e.g. the SHEBA (Surface Heat Budget of the Arctic Ocean, Perovich et al., 1999) experiment and AOE-2001 (Arctic Ocean Experiment 2001, Leck et al., 2004; Tjernström et al., 2004; Tjernström, 2005).

Previous modelling studies of the Arctic climate system have focussed on special regions, for example the marginal ice zone (Vihma et al., 2003), on shorter periods (Rinke et al., 1999, 2000), or have used single column models (Pinto et al., 1999). Only a few studies cover larger areas (Rinke et al., 2000, 2003) or longer time periods (e.g. Christensen and Kuhry, 2000), and systematic model evaluations are rare. The Arctic regional climate model intercomparison project (ARCMIP, Curry and Lynch, 2002, http://curry.eas.gatech.edu/ARCMIP/index.html) aims at identifying model deficiencies and at improving the description of Arctic climate processes in numerical models. This is achieved by carrying out controlled regional-model experiments for the Arctic. An underlying strategy is to use regional models to improve global climate modelling. In a regional model, the larger-scale climate can be controlled, by prescribing the lateral boundary from global analyses. Remaining systematic errors in the regional models are then likely related to deficiencies in their description of sub grid-scale processes – the parameterisations. Problems become more easily isolated, and can be dealt with more easily, than within the framework of global model control experiments. We can also afford to operate regional models today at resolutions expected in future GCMs. The higher spatial resolution in a regional model also allows a better representation of important feedback processes.

In ARCMIP several models are intercompared and compared with observations. All models are operated in the same way, and the first ARCMIP experiment is a 13-month long simulation for the western Arctic, from September 1997 through September 1998. In this paper, we evaluate boundary-layer results at the SHEBA column from six such models. From analysing the model errors we are able to evaluate the skill of these state-of-the-art models in simulating the Arctic boundary layer and also hope to learn how to improve the models. In Section 2, the experiment and the different models are presented. While the focus is on the boundary layer, this is interpreted in a broad sense. Results on some surface properties are discussed in Section 3, for example for the energy fluxes at the surface, which are critical to the strong ice/albedo feedback. The vertical structure of the simulated lower troposphere is compared to SHEBA soundings in Section 4. A summary discussion is found in Section 5.

2. The Model Experiment

2.1. Experiment set-up

The six models in Table I were all set up on a common model domain over the western Arctic covering approximately 3500×2750 km^2 (Figure 1), and determined by the SHEBA ice-drift track. In the south, it covers most of Alaska, the Bering Strait and north-eastern Siberia and to the north reaches into the pack ice to about 85° N (see the ARCMIP home page for exact specifications). The models have slightly different grid systems, and the grid points, in general, do not coincide exactly. The target resolution of about 50 km is however roughly the same in all of the models. As an example, the grid points shown in Figure 1 are taken from COAMPS$^{\text{TM}}$. Vertical resolutions and time steps were also different in the different models; see Table I. The lateral boundary forcing was provided at 6-h intervals using ECMWF (European Centre for Medium Range Weather Forecasts) operational analyses, the same for all models. Sea-surface temperature (SST) and ice fraction were also prescribed the same for all models, taken from AVHRR (Advanced Very High Resolution Radiometer) and SSMI (Special Sensor Microwave Imager) satellite observations, respectively, see the ARCMIP home page. The surface temperature over land, however, was derived from each model's surface energy balance calculations. To isolate the atmospheric model problems from oceanic and cryospheric problems, the ice-surface temperatures were also prescribed from AVHRR data, at a 6-h resolution.

All model results are compared with measurements from the SHEBA Atmospheric Surface Flux Group (ASFG) instrumented tower (Persson et al., 2002) and with data from radiosoundings performed through the whole year at the SHEBA site. For all of the comparisons, we have used model output from the grid point that is closest to the SHEBA track (Figure 1). It is worth emphasizing that all simulations were run continuously through the whole year, without the benefit of data assimilation. They are forced only by the specifications of the boundary conditions, thus allowing systematic errors to grow. However, the analyses on the lateral boundaries from ECMWF are taken from a data assimilation cycle, in which soundings and surface observations from SHEBA were ingested. Statistics for November 1997 to January 1998 suggest that roughly 85% of the soundings

TABLE I
Summary of the participating models.

Model name: Name of responsible group	Vertical grid system Total # / # below 500 m	 Lowest level (m)	 Type	Time step (min)	Surface layer scheme	PBL scheme	Main reference
ARCSyM: University of Colorado	23/4	35–40	P	2.5	Mellor-Yamada Level 2	Mellor-Yamada Level 2.5	Lynch et al. (1995)
COAMPS^{TM}: Stockholm University	30/7	15	Z	1.5	Louis	Mellor-Yamada Level 2.5	Hodur (1997)
HIRHAM: Alfed-Wegener Institute	19/3	25–30	P	5	Louis	Level 1.5	Christensen et al. (1996)
Polar MM5: University of Colorado	23/4	30-50	P	2.5	Mellor-Yamada Level 2	Mellor-Yamada Level 2.5	Cassano et al (2001)
RCA: Swedish Meteorological and Hydrological Institute	24/3	70–85	P	30	Louis	CBR	Jones et al. (2004)
REMO Max-Planck Institute for Meteorology	20/3	55–65	P	5	Louis	Level 1.5	Jacob (2001)

In the vertical grid system column (type),“Z” means that a geometric system is used and “P” means that a system based on pressure is used. In both cases, different scaling may have been applied to account for terrain height.

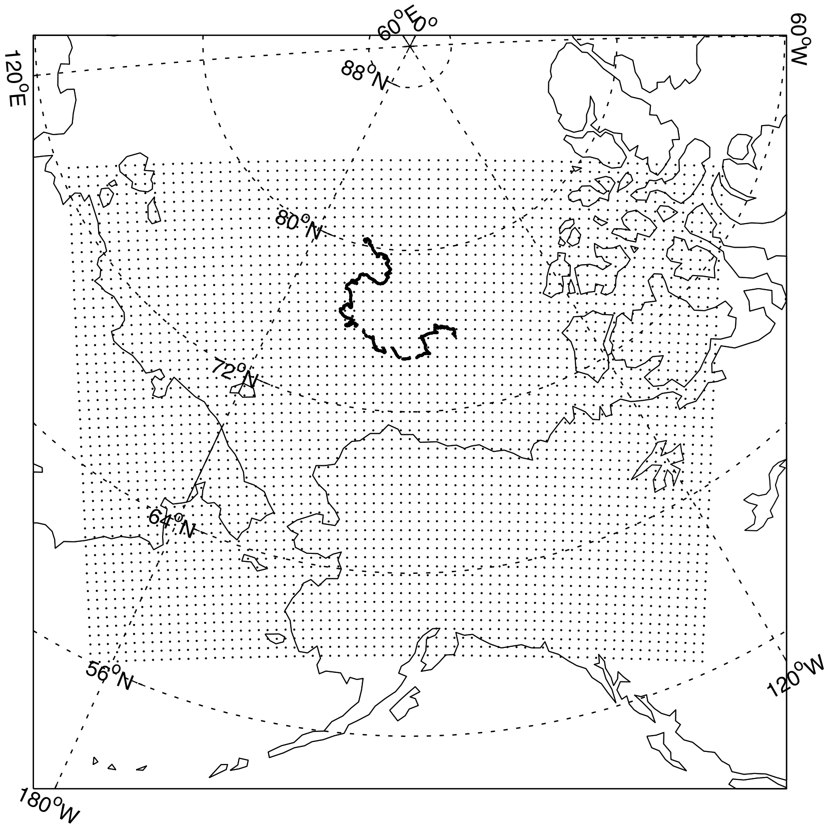

Figure 1. The geography of the model domain. The dots represent the computational grid of the COAMPS^{TM} model and the path north of Alaska is the SHEBA ice drift track, starting at the south-east point and ending at the north-west point.

reached ECMWF and entered the analysis (C.A Bretherton, personal communication, 2000).

2.2. The models

The regional-scale models included in this study are summarised in Table I: ARCSyM (Arctic Regional Climate System Model, Lynch et al., 1995); COAMPS^{TM} (Coupled Ocean-Atmosphere Mesoscale Prediction System, Hodur, 1997); HIRHAM4 (HIRLAM – High Resolution Limited Area Model with physics from ECHAM4, a GCM based on ECMWF forecast models modified and extended in HAMburg, Christensen et al., 1996); Polar MM5 (Polar version of NCAR/Pennsylvania State University fifth generation Mesoscale Model, Bromwich et al., 2001; Cassano et al., 2001); REMO (REgional MOdel from the Max Planck Institute, Jacob, 2001); and RCA (Rossby Centre Atmospheric model, Jones et al., 2004).

There are similarities and differences between the models, but they all have different grid architecture, horizontal as well as vertical. The main differences are in the vertical grid, with models using either pressure (or scaled pressure) or geometrical height (or scaled geometrical height) as a vertical coordinate. In the latter, the heights of a grid point above the surface are fixed in time, while in the former they vary. While all of the models have more vertical levels than typical GCMs, both near the surface and in total (Table I), there are also differences in resolution, in particular near the surface and the number of points below 500 m, between 3 and 7, is still low. The height of the lowest level varies significantly, from 15 m to about 80 m above ground level.

All of the models have different model physics and many have a more sophisticated boundary-layer parameterisation than that common in many current GCMs, although GCM development is intensive and this is changing rapidly. All have planetary boundary-layer (PBL) schemes that are based on a prognostic turbulent kinetic energy (TKE), sometimes somewhat loosely referred to as second-order closure. Although appearing superficially similar, there are two basic types of TKE scheme. One, often referred to as 'Mellor-Yamada Level 2.5', is based on a systematic scale analysis of the full set of equations for all ensemble-averaged second-order turbulent moments (Mellor and Yamada, 1974). After justified simplifications and an inversion of the remaining equation-system matrix, turbulent eddy-exchange coefficients are obtained as complex functions of the TKE, a length scale and the vertical gradients of virtual potential temperature and wind speed. This can be thought of as a top-down process. The other approach (e.g. Brinkop and Roeckner, 1995) is more bottom-up. From dimensional arguments one can argue that eddy-exchange coefficients must be proportional to a (mixing) length times the square root of the TKE. This closure is sometimes referred to as 'Level 1.5'; the proportionality can be either a constant or a semi-empirical stability-dependent analytical function. To a first-order approximation, the top-down method approaches the same functional relationship as the bottom-up approach. The 'Mellor-Yamada Level 2.5' is used in ARCSyM, COAMPS$^{\mathrm{TM}}$ and Polar MM5, while a 'Level 1.5' approach is used in HIRHAM and REMO. RCA uses the so-called CBR scheme (CBR after the authors of Cuxard et al., 2000), a method somewhat similar to 'Mellor-Yamada Level 2.5' but applying this only for dry variables, and thus not allowing condensation to have any effect on buoyancy.

Both types of closure rely on a length-scale formulation. In a sense, some of the semi-empirical aspects of the problem is shifted from prescribing the vertical shape of eddy-exchange coefficients, so-called first-order closure often used in GCMs, to prescribing the functional shape of a length scale. As the eddy-exchange coefficient additionally depends on TKE, a forecast variable, the closure assumptions are removed one step farther from the solution and the eddy-exchange coefficients are thus allowed to respond to the local dynamics of the flow. Such schemes are sometimes called 'non-local'. It is important, however, to realise this usually implies that a 'correction' is imposed so that the flux of a property can be directed opposite to its gradient, and does not mean that the scheme is truly non-local. At a basic level, all ensemble-average closure models are local in the sense that a flux is to a first order determined by a local vertical gradient. One way to attempt to mimic non-locality is through the mixing length. In the CBR scheme this is accomplished by determining the mixing length from vertically integrated properties.

Even the most sophisticated PBL schemes have to be provided boundary conditions at the surface – the surface fluxes. Surface-layer descriptions in the majority of the models (COAMPS$^{\text{TM}}$, HIRHAM, RCA and REMO) utilise the Louis scheme (Louis, 1979), which is based on Monin-Obukhov surface-layer similarity theory (MOST) and is widely used in weather forecast models. The non-linear surface-layer theory is simplified by fitting polynomials to the so-called ψ functions from MOST, retaining some of the similarity while allowing a faster and more flexible system. The polynomials are, for example, slightly adjusted so that they allow for mixing at super-critical Richardson numbers. ARCSyM and Polar MM5 uses the 'Mellor-Yamada Level 2' scheme for the surface fluxes (Mellor and Yamada, 1982), which is an analytical simplification of the 'Level 2.5' scheme. Although it is not directly based on MOST, it has been shown to reproduce almost the same features. While the surface-layer scheme uses local surface-layer mean gradients to calculate surface fluxes, it is important to realise that these gradients in turn are dependent on the PBL scheme, and *vice versa* – neither works in isolation. Moreover, since MOST assumes stationarity, the application of these schemes implies that the TKE equation, which allows for non-stationarity, is continuously forced to a stationary solution.

2.3. The error analysis

Statistical results for the annual cycle are summarised in Tables II to IV, based on instantaneous three-hourly data. A limiting factor for the statistics is the available number of observations, which differs for different variables; this number is given in the sub-heading in each Table. The results in the figures are based on either daily or weekly averages, for clarity. When the observational coverage within a time interval was $< 50\%$, that time interval is not used. Errors are defined as model minus observation. In addition to standard error measures (mean bias, root-mean-square error and correlation

TABLE II

Annual error statistics, for the 2-m temperature (K), low-level specific humidity (g kg^{-1}) and 10-m wind speed (m s^{-1}), calculated from 3-hourly data.

2-m temperature $N = 1828$; $\sigma_o = 12.9$ K	Bias (K)	σ_m (K)	RMSE (K)	R	IoA
ARCSyM	0.08	11.0	3.77	0.96	0.97
COAMPS^{TM}	−0.17	11.7	3.46	0.96	0.98
HIRHAM	0.20	11.3	3.37	0.97	0.98
PMM5	0.41	11.6	3.42	0.97	0.98
RCA	−0.20	11.4	3.25	0.97	0.98
REMO	−0.28	11.6	3.25	0.97	0.98
Humidity (2-m/lowest level) $N = 2456$; $\sigma_o = 1.41$ g kg^{-1}	Bias (g kg^{-1})	σ_m (g kg^{-1})	RMSE (g kg^{-1})		
ARCSyM about 35–40 m	−0.27	1.10	0.50	0.97	0.96
COAMPS^{TM} 15 m	−0.01	1.22	0.35	0.97	0.98
HIRHAM 2 m	−0.19	1.10	0.45	0.98	0.97
PMM5 about 40 m	−0.01	1.26	0.34	0.97	0.98
RCA 2 m	−0.24	1.16	0.44	0.98	0.97
REMO 2 m	−0.16	1.19	0.44	0.96	0.97
10-m wind speed $N = 2258$; $\sigma_o = 2.55$ m s^{-1}	Bias (m s^{-1})	σ_m (m s^{-1})	RMSE (m s^{-1})		
ARCSyM	−0.06	2.20	2.09	0.62	0.78
COAMPS^{TM}	0.38	2.47	2.04	0.68	0.82
HIRHAM	0.47	2.81	2.29	0.65	0.80
PMM5	1.47	2.99	2.81	0.64	0.75
RCA	−0.98	2.17	2.05	0.72	0.81
REMO	0.30	2.84	2.07	0.72	0.84

Note that for humidity, some models use the 2-m value while others use the lowest model grid point. In each subheading, n is the total number of observations used for each variable, while σ_o is its standard deviation. 'Bias' is the mean error, σ_m is the standard deviations of each models and 'RMSE' is the root-mean-square-error. R is the correlation coefficient and IoA is the 'Index of Agreement'.

TABLE III
Same as Table II, but for the four components of the radiation heat flux (W m^{-2}), see each subheading.

Downward shortwave radiation $N = 1500$; $\sigma_o = 155$ W m^{-2}	Bias (W m^{-2})	σ_m (W m^{-2})	RMSE (W m^{-2})	R	IoA
ARCSyM	−4.3	150	74.6	0.88	0.94
COAMPS^{TM}	−11.1	140	64.0	0.91	0.95
HIRHAM	−22.4	153	112.0	0.75	0.86
PMM5	11.3	171	94.0	0.84	0.91
RCA	15.8	156	86.0	0.85	0.92
REMO	−41.0	146	102.6	0.81	0.88
Upward shortwave radiation $N = 1461$; $\sigma_o = 112$ W m^{-2}					
ARCSyM	8.8	121	63.4	0.86	0.92
COAMPS^{TM}	−10.7	102	50.9	0.90	0.94
HIRHAM	−36.8	96	83.5	0.75	0.83
PMM5	21.0	136	80.2	0.82	0.89
RCA	25.1	123	71.6	0.84	0.90
REMO	−49.7	92	85.3	0.79	0.83
Downward longwave radiation $N = 2487$; $\sigma_o = 62.7$ W m^{-2}					
ARCSyM	−9.9	68.4	34.9	0.87	0.93
COAMPS^{TM}	−18.3	65.0	42.1	0.83	0.89
HIRHAM	−10.0	57.2	29.8	0.89	0.94
PMM5	−34.7	60.7	53.2	0.79	0.82
RCA	−0.3	51.8	29.6	0.88	0.93
REMO	−4.7	60.6	29.1	0.89	0.94

Upward longwave radiation $N = 2427$; $\sigma_o = 51.6$ W m^{-2}					
ARCSyM	−1.6	43.9	15.6	0.96	0.97
COAMPS^{TM}	−3.6	49.4	15.1	0.96	0.98
HIRHAM	−1.3	43.9	14.2	0.97	0.98
PMM5	−0.9	45.4	14.1	0.97	0.98
RCA	−0.8	44.8	14.1	0.97	0.98
REMO	−1.4	45.0	13.9	0.97	0.98

Note that the annual error statistics for shortwave radiation is only calculated for cases when the downward flux is > 1 W m^{-2} were used.

TABLE IV
Same as Table II, but for the turbulent surface fluxes of sensible and latent heat (W m^{-2}) and for the turbulent surface friction velocity (m s^{-1}).

Turbulent sensible heat flux $N = 2129$; $\sigma_o = 8.9$ W m^{-2}	Bias (W m^{-2})	σ_m (W m^{-2})	RMSE (W m^{-2})	R	IoA
ARCSyM	−1.0	39.7	39.8	0.10	0.22
COAMPS^{TM}	−2.0	12.8	12.7	0.39	0.60
HIRHAM	−2.8	12.9	14.1	0.23	0.48
PMM5	−2.7	13.2	15.3	0.11	0.39
RCA	−4.1	20.1	20.6	0.21	0.41
REMO	1.3	13.0	14.5	0.17	0.45
Turbulent latent heat flux $N = 1188$; $\sigma_o = 1.9$ W m^{-2}					
ARCSyM	6.1	8.8	10.4	0.31	0.22
COAMPS^{TM}	3.5	5.3	12.7	0.46	0.60
HIRHAM	1.7	5.5	5.6	0.24	0.33
PMM5	0.3	4.3	4.5	0.12	0.32
RCA	4.7	10.6	11.2	0.30	0.22
REMO	1.6	5.5	5.5	0.27	0.36
Turbulent friction velocity $N = 2008$; $\sigma_o = 0.12$ m s^{-1}	Bias (m s^{-1})	σ_m (m s^{-1})	RMSE (m s^{-1})		
ARCSyM	0.080	0.15	0.14	0.61	0.71
COAMPS^{TM}	0.048	0.13	0.11	0.68	0.79
HIRHAM	0.037	0.12	0.11	0.63	0.77
PMM5	−0.004	0.13	0.11	0.60	0.77
RCA	0.117	0.18	0.18	0.66	0.67
REMO	0.061	0.15	0.13	0.67	0.76

coefficient), we also make use of the ‘Index of Agreement’, IoA*. This can be considered as an alternative correlation coefficient that takes into account phase differences between the compared signals. As an example, the correlation coefficient between two sine functions a quarter of a wavelength out of phase is zero. The IoA is about 0.4, for the same amplitude, and so similarity can be detected even in poorly correlated signals.

Any modelling-error analysis is incomplete without considering possible measurement problems. First, there is an inherent discrepancy between modelled grid-point averages and true point measurements. Using data from a whole year, some of the heterogeneity problem may hopefully average out, but for some variables it may contribute to a systematic error. For example, the models account for fractional ice within a 50×50 km^2 grid box while point measurements generally do not, since the measurements have a limited fetch and are located on more or less solid ice. For latent heat, the presence of open water in a grid box will lead to values larger than (or equal to) those represented by the measurements near the surface on a large ice sheet. Open water during winter will similarly always lead to a larger sensible heat flux than for homogeneous ice conditions, but the opposite will never happen. Even with an extensive dataset, such as that from SHEBA, there is not much to be done about this problem, other than to be aware of it in the analysis.

Second, there are always problems with measurements, and these problems differ for the different variables. Persson et al. (2002) provide a summary of the SHEBA ASFG measurements. For near-surface wind-speed and temperature, easily maintained high-quality instruments are available and therefore such data usually have both high quality and high recovery rates. Atmospheric humidity is significantly more difficult to measure, in particular in cold conditions. The mean humidity measurements at SHEBA were performed with relative-humidity sensors, converting to specific humidity using temperature and pressure. The accuracy of this type of instrument at high relative humidity is problematic (Persson et al., 2002) and humidity measurements are less reliable than temperature, for example.

Radiation measurements at the ASFG site were made with standard sensors of good quality and also have a high recovery rate, but the largest problem in a model-comparing context is their representativity. For example, the upward shortwave radiation measurements are strongly influenced by the local albedo of a small patch of surface beneath the instrument, while the upward flux in a model is an area average. During Arctic summer, the distribution of melt ponds and open water can allow a spot-measured albedo to be an overestimate (Intrieri et al., 2002). The ASFG radiation measurements were affected by a near-by melt pond (Persson et al., 2000), but were likely

* The IoA is defined as $\mathrm{IOA} = 1 - \frac{\sum_1^n (P-O)^2}{\sum_1^n (|P-\bar{O}|+|O-\bar{O}|)^2}$, where P and O are predicted and observed values, respectively, n is the number of observations and an overbar indicates a time average.

less affected by more typical melt-pond conditions due to the need to have the site located on solid ice. Additional albedo measurements were made along a line with a more representative mix of melt ponds, leads and snow/ice covering a shorter time period.

Although routine long-term turbulence measurements have become increasingly possible since the introduction and refinement of sonic anemometers, this type of measurement will always require careful screening. In particular the heat fluxes can be problematic. The sensible heat flux is derived from measurements of the speed-of-sound, converted to the so-called 'sonic temperature' (close to the virtual temperature); this is often noisy and problematic. Also, direct measurements of the latent heat flux require very fast and accurate humidity measurements, while even long-term stable mean humidity is difficult to measure. SHEBA turbulence measurements were carefully quality controlled (Persson et al., 2002).

Sounding equipment makes use of inexpensive instruments for temperature and humidity measurements and are less reliable, typically ±0.5 °C for temperature and ±5 % for relative humidity. Wind measurements on sounding systems are based on the principle of tracking the horizontal motions of the sonde. In SHEBA, a Global Positioning System sensor was used for this purpose and its accuracy is typically ±0.5 m s^{-1}. As the senso requires reception of signals from a number of satellites, the wind measurements are sometimes missing and winds close to the surface are always questionable. High wind speeds, above about 30 m s $^{-1}$, are very often missing in the SHEBA sounding dataset. Thus, although temperature is probably reasonably good from the soundings, moisture is often of more doubtful quality, and winds are always a problem.

Finally, interpreting model performance objectively is not straightforward, and we have adopted the principles outlined by Hanna (1994). A good result is thus signified by a small bias, similar standard deviations of the model result and the corresponding observation, a root-mean-square error that is smaller than both of these, and finally a high correlation coefficient and/or IoA.

3. Near-surface Results

3.1. Forcing

Prescribing the temperature of ocean and sea-ice surfaces in all models using independent measurements constrains the modelled 2-m temperatures, and other parameters relying strongly on it, to be close to the corresponding measurements, provided that the surface temperatures retrieved from the AVHRR measurements are sufficiently accurate. The surface temperature was here retrieved using the CASPR (Cloud and Surface Parameter Retrieval) algorithm (Key, 2002); this utilises primarily AVHRR radiative temperatures during clear-sky conditions, and empirical interpolative relationships to estimate the surface temperatures during cloudy conditions, to obtain 50×50 km^2 averaged surface temperatures at a six-h intervals. The root-mean-square error estimates for this dataset are 1–2 K for clear-sky conditions, but substantially larger, around 6 K, for cloudy conditions (Key, 2002). Although the annual cycle (not shown) agrees well with the observed ASFG surface temperature from SHEBA, there are significant systematic errors that become evident when the error is plotted as a function of the latter (Figure 2). The CASPR surface temperature is higher than that from the ASFG during cold winter conditions, on average by about 5 °C for temperatures below −30 °C, and slightly lower, about 2 °C, for temperatures higher than −20 °C. The scatter around the mean error is significant, with most of the scatter due to the higher variability in the SHEBA surface temperature.

The differences between the CASPR and SHEBA surface temperatures shown in Figure 2 are likely a combination of errors in the CASPR data and differences due to the spatial averaging inherent in the satellite data; this illustrates clearly the problem of comparing single point observations with

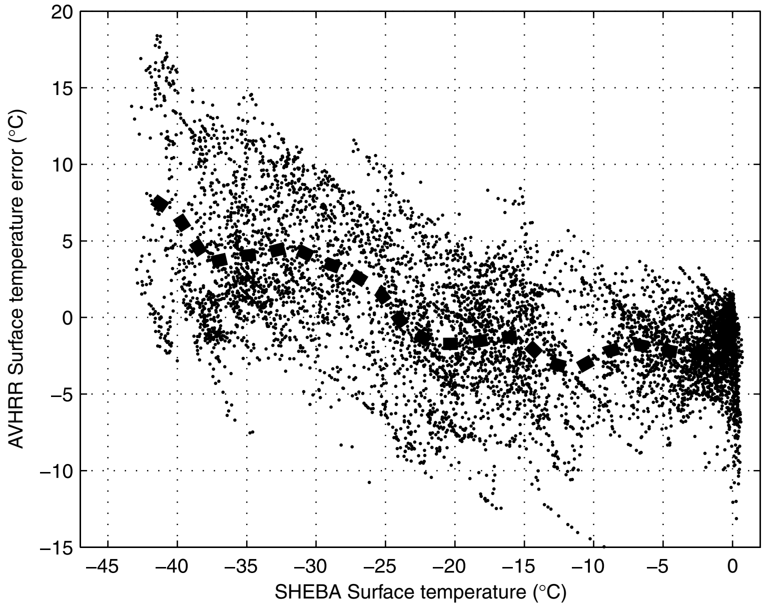

Figure 2. Scatter plot of the AVHRR surface temperature error, defined as the AVHRR temperature minus the measured SHEBA surface temperature, plotted against the latter (both in °C, dots). The thick dashed line represents the average taken over 1 °C intervals of the measured surface temperature.

area-averaged measurements. Overland et al. (2000) discuss the effects of snow and/or ice thickness variations and show surface temperature differences of the order of 10 °C in a 100 × 100 km^2 box around the SHEBA site. The systematic warm bias during the lowest temperatures may be due to problems in the CASPR routine during cloudy conditions. Changes in the cloud cover also cause very rapid changes in the surface temperature that are not resolved by the six-hourly analyses here. The annually averaged root-mean-square difference in Figure 2 is about 4 K, which is smaller than the estimated error during cloudy conditions but larger than that for clear sky conditions by about a factor of two; it increases to about 5.5 K for temperatures < -30 °C. Considering the cloud fractions observed during SHEBA (Persson et al., 2002), these errors are consistent with the error estimates in Key (2002).

The wisdom of prescribing the ice-surface temperature may thus be debated. One could view this as a 'best case scenario', and surface temperature errors would probably be larger if the ice-surface temperature had been modelled, due to the complexities involved in modelling the evolution of the snow and ice cover, and the fraction of melt ponds. However, once that decision was taken there is a limited choice on the data to use. If one wants to be free of the model biases appearing in all analyses in data sparse regions, only satellite data remain. Distinguishing between the surface and clouds in the Arctic remains a problem. The CASPR algorithm used here was evaluated by Liu et al. (2005), and was found to have a significantly smaller bias, both on an annual average and annual cycle basis, than available re-analysis data and ISCCP (International Cloud Satellite Climatology Project) data. Consequently, the modelled near-surface air temperature and other variables closely related to the surface temperature should thus be reasonably close to the observations, within the errors indicated in Figure 2. Errors in, for example, wind speed, radiative and turbulent fluxes and low-level vertical structure should, however, predominantly reflect real deficiencies in the different models.

The ensemble-averaged simulated mean sea-level pressure follows the measurements closely on average and the inter-model spread is quite small (Figure 3). This indicates that the model domain is small enough that the regional models are well constrained on the synoptic scale by the ECMWF analyses used as lateral boundary conditions. This is important, since it means that we can assume that the 'larger-scale weather' is essentially correct and can thus proceed to investigate more locally determined processes and variables that represent the individual model physical parameterisations.

3.2. Near-surface mean parameters

Figure 4 shows time series of the simulated 2-m temperature error from the individual models and the corresponding temperature measured on the

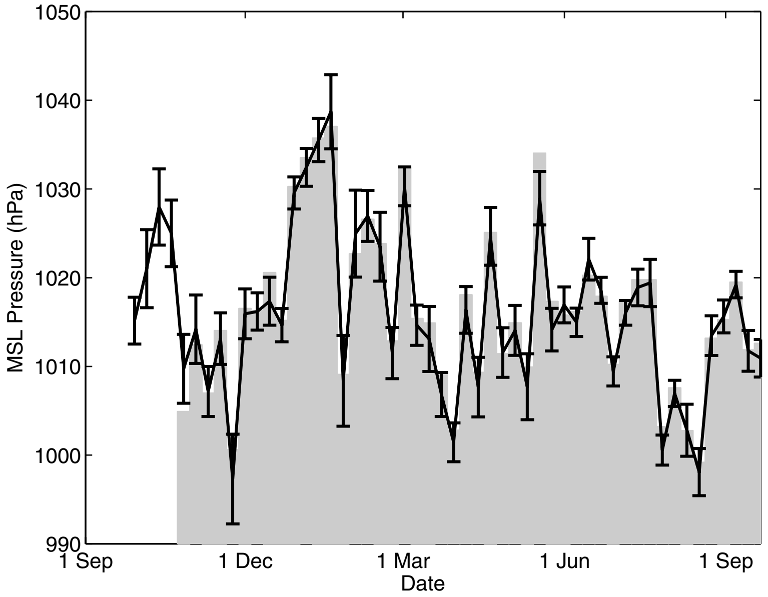

Figure 3. Plots of year long time series of weekly averaged model-mean mean sea-level pressure (hPa) from the six models (black line), the inter-model standard deviation (error bars) and the SHEBA observations (grey bars). The average for a particular week of SHEBA observations is missing if less than 50% of the three-hourly observations are available.

ASFG mast at the SHEBA site, for two periods. Table II presents the corresponding annual error statistics. Note, that the 2-m temperature does not conform to any grid level in the models, and is a standard interpolation result that should be consistent with the surface-layer scheme of each model. As expected, the biases are quite small on an annual scale and the root-mean-square (RMS) errors are about a third of the standard deviations from either models or observations, which are similar. The annual biases range from about −0.3 °C in REMO to about 0.4 °C in Polar MM5, while the correlation coefficients and the IoA are close to unity.

During some, but not all, winter periods when the observed temperature is < −30 °C (around 240 K or less), there is a tendency in some models to be too warm. This winter error is consistent with the AVHRR temperature error (Figure 2). During less cold periods of winter and during spring, all of the models are similar. In summer the models split into two groups, with two models (Polar MM5 and COAMPS$^{\mathrm{TM}}$) having near-surface temperatures close to 0 °C much of the time, while the rest (ARCSyM, HIRHAM, RCA and REMO) have a lower temperature, about −2 °C. Polar MM5 and sometimes REMO show a tendency to swap between these two states, at least early in the summer. Some models have periods with much lower

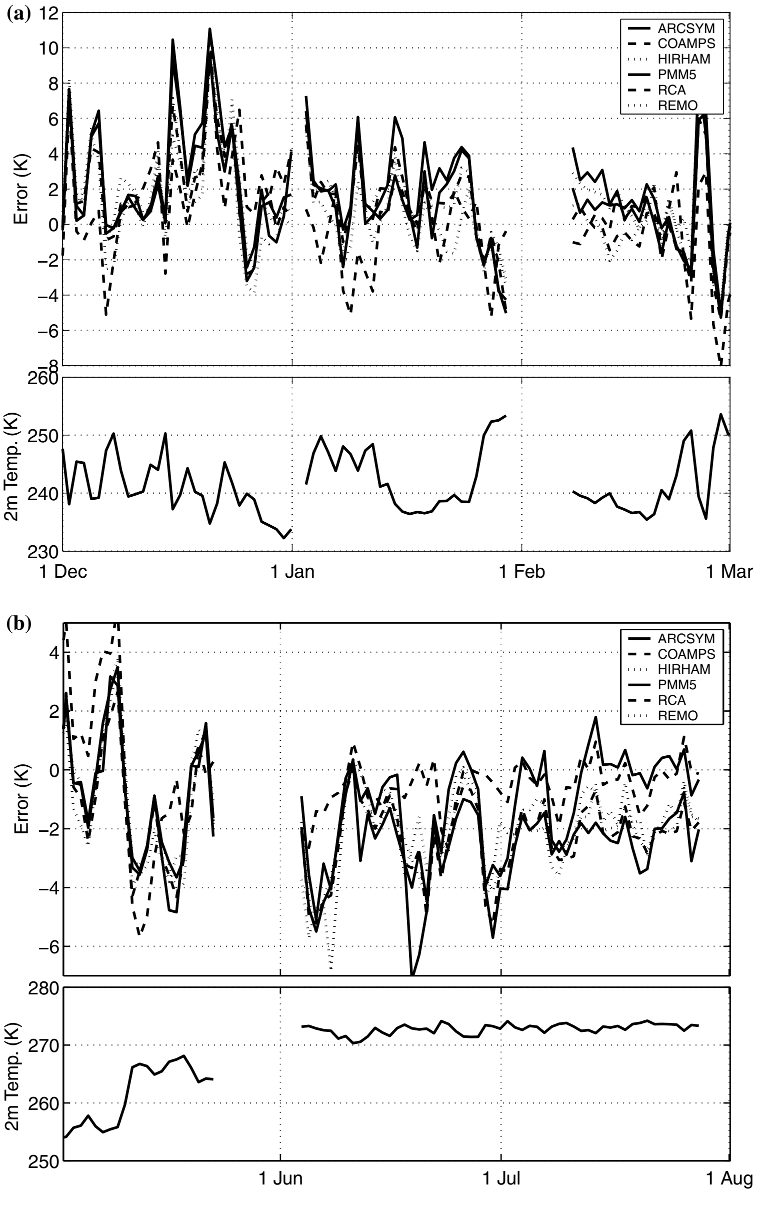

Figure 4. The time evolution of the (lower panels) daily averaged observed temperature and (upper panels) modelled temperature error at 2 m (K) for (a) winter and (b) late spring and early summer. The same legend is used throughout this paper for similar plots.

temperatures in early summer, briefly down to about −8 °C, in particular ARCSyM. This is consistent with the prescribed satellite temperatures but cannot be verified by the measurements.

Summer near-surface temperatures are effectively constrained by melting and freezing at the surface (e.g. Tjernström, 2005) and should be constrained roughly between −2 and 0 °C. It seems that some models consistently "want to warm", while other models "want to cool", outside this interval. While the first two models adhere to the melt temperature of fresh water, around 0 °C, the rest adhere to the melt temperature of salty ocean water, near −2 °C. The measurements indicate that the former is more accurate. In the ARCMIP procedure, the surface temperature is affected by the amount of open water, where the SST was set to −2 °C as long as there is both ice and open water present. There is also an upper limit of 0 °C for the ice surface. Different models handle the combination of partial ice cover and the prescribed AVHRR surface temperature differently on a technical level. Thus, even with one unified surface temperature, with these additional constraints there are differences between the actual surface temperatures used in the models, ranging from less than 1 °C in summer to 1 or 2 °C in winter. This may appear as an artificial difference. Note, however, that the same technical differences would appear even if the models had been allowed to calculate their own ice-surface temperatures through, for example, a thermodynamic ice model.

Time series of the specific humidity error from the individual models and the SHEBA humidity measurements are shown in Figure 5; note that here the 10-m observation level is used. The data recovery from SHEBA is much higher for the 10-m than for the 2-m level, and the two differ insignificantly. Moreover, not all of the models provided a 2-m value as a standard output. As a compromise, we compare the 10-m SHEBA humidity to the 2-m humidity from HIRHAM, RCA and REMO, while for COAMPS^{TM}, ARCSyM and Polar MM5, the lowest model level is used; see Table 2. The annual RMS error is again small, ≤ 0.5 g kg^{-1}, which is 30 to 50% of the standard deviations. All models are biased dry, from about -0.5 g kg^{-1} (for ARCSyM) to nearly zero (for Polar MM5 and COAMPS^{TM}). However, the annual errors are dominated by the summer conditions when the humidity is an order of magnitude larger than in winter.

To a first order, the differences between the models in near-surface temperature explain the differences in low-level specific humidity. Note that near-surface specific humidity is always close to saturation with respect to ice (Andreas et al., 2002). During summer, the two models with the most accurate surface temperature (Polar MM5 and COAMPS^{TM}) also have the smallest moisture error. During winter, the absolute value of the specific humidity is very low so that errors in temperature have a very small absolute effect on the humidity error. During the periods with the very lowest temperatures, the highest humidity is not found in the warmest model. Instead there are three groups of models, with the highest winter moisture found in the Polar MM5. COAMPS^{TM} and HIRHAM have moisture values that are

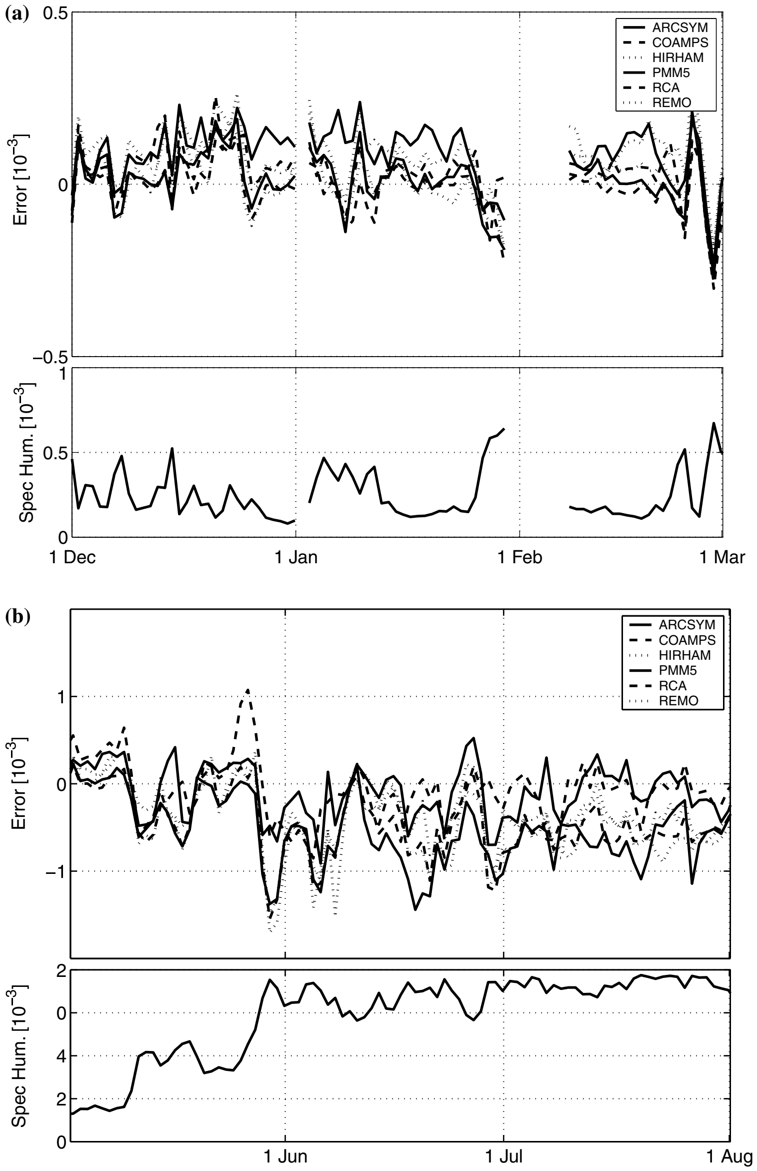

Figure 5. Same as Figure 4, but for the near-surface specific humidity (g kg^{-1}). See the text for discussion on the heights for the model results. The legend is the same as in Figure 4: ARCSyM, COAMPS^{TM} and HIRHAM are black while Polar MM5, RCA and REMO are grey, solid, dashed and dotted, respectively.

slightly high while ARCSyM, REMO, and RCA are quite close to the observations. A systematic difference in near-surface specific humidity of about 0.5 g kg^{-1} in summer could easily generate systematic differences in low-level cloud cover, a quasi-persistent feature in the summer, while the effects of errors in the wintertime humidity on cloud cover are probably smaller. First, the very low winter temperatures result in a low specific humidity. Second, since the near-surface humidity is close to ice saturation, the relative humidity with respect to liquid water is kept lower through the winter and the impact on liquid-water cloud formation by a near-surface moisture error is therefore likely smaller.

While the prescribed surface temperature constrains the near-surface air temperature and humidity, the simulated 10-m wind speed (Figure 6 and Table 2) is determined by a combination of synoptic-scale dynamics and the turbulent momentum flux. All models follow the temporal variability in the observations quite closely. However, the different models have quite different wind-speed biases. Polar MM5 systematically has the highest wind speeds, for an annual average about 1.5 m s^{-1} too high, while RCA winds are the lowest, on average about 1 m s^{-1} too low. In between these, the inter-model differences are smaller. The winds are somewhat too high also in HIRHAM, REMO and COAMPS$^{\mathrm{TM}}$ (by about 0.5 m s^{-1} or less on an annual average) while ARCSyM has the smallest annual bias, near zero. In some models, the bias is also a slight function of wind speed itself. The biases are thus larger for higher winds speeds in RCA (at >10 m s^{-1}) and in Polar MM5 (at > 5 m s^{-1}), see Figure 7. Still, the annual RMS errors are of the order of 2–3 m s^{-1} and the correlation coefficients are around 0.7, highest in RCA and REMO, while the IoA is about 0.8, highest in COAMPS$^{\mathrm{TM}}$ and lowest in Polar MM5. These results are surprisingly good and it is not until we examine the turbulent momentum fluxes that we find hints of more systematic problems (see below).

3.3. Surface radiative fluxes

Figure 8 shows three components of the surface radiation fluxes and the albedo from the SHEBA observations, and the corresponding model errors. Considering first the incoming shortwave radiation (Figure 8a; Table 3), this is largely determined by the clouds while the surface albedo also plays a secondary role, due to multiple reflections between the snow surface and the clouds. On an annual scale, all models perform reasonably well. The largest error is found in REMO, followed by HIRHAM, at about −40 and −20 W m^{-2}, respectively while ARCSyM is closest to the measurements with a −4 W m^{-2} bias. COAMPS$^{\mathrm{TM}}$, Polar MM5 and RCA have positive biases around 10–15 W m^{-2}. All of the models appear to capture the incoming shortwave radiation during spring and early summer reasonably well, in particular in early spring and late summer into autumn. During the early

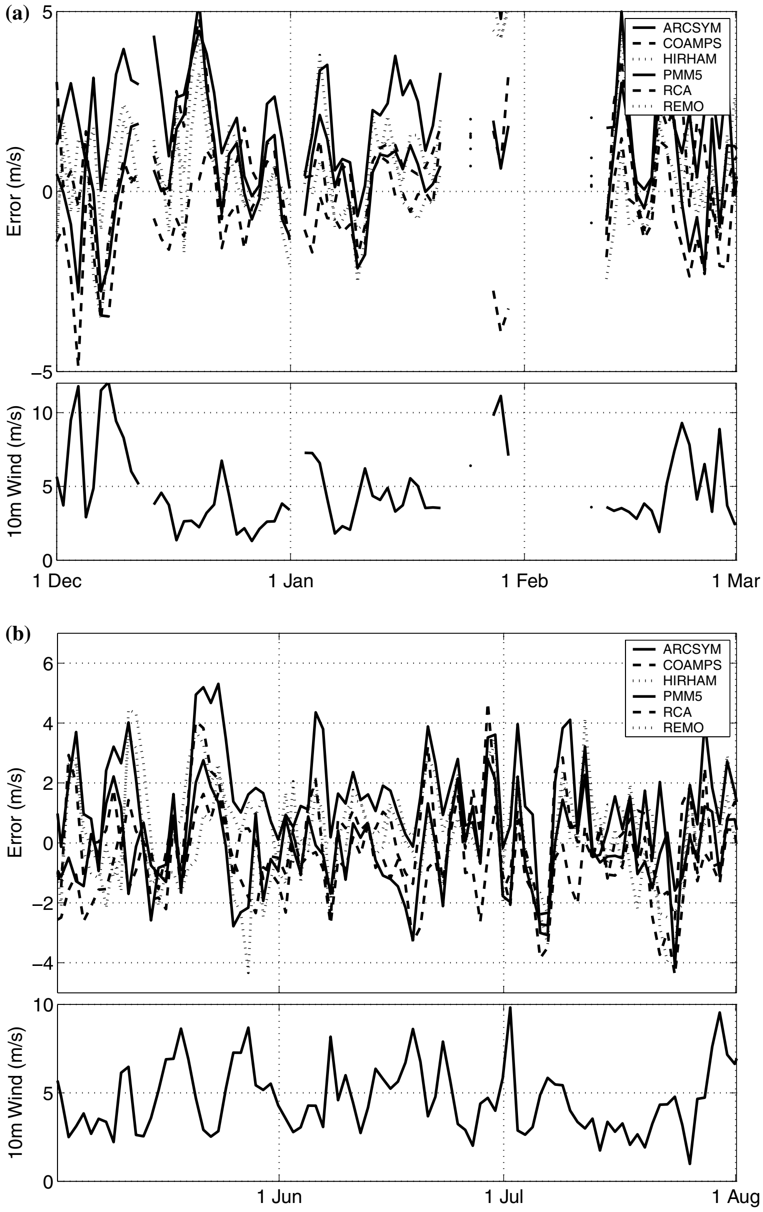

Figure 6. Same as Figure 4, but for the wind speed at 10 m (m s^{-1}). The legend is the same as in Figure 4: ARCSyM, COAMPS^{TM} and HIRHAM are black while Polar MM5, RCA and REMO are grey, solid, dashed and dotted, respectively.

summer, there are systematic differences, with ARCSyM, RCA and Polar MM5 agreeing best with the measurements, while the three remaining models are low from May to July. Representing the fluxes in an accumulative sense,

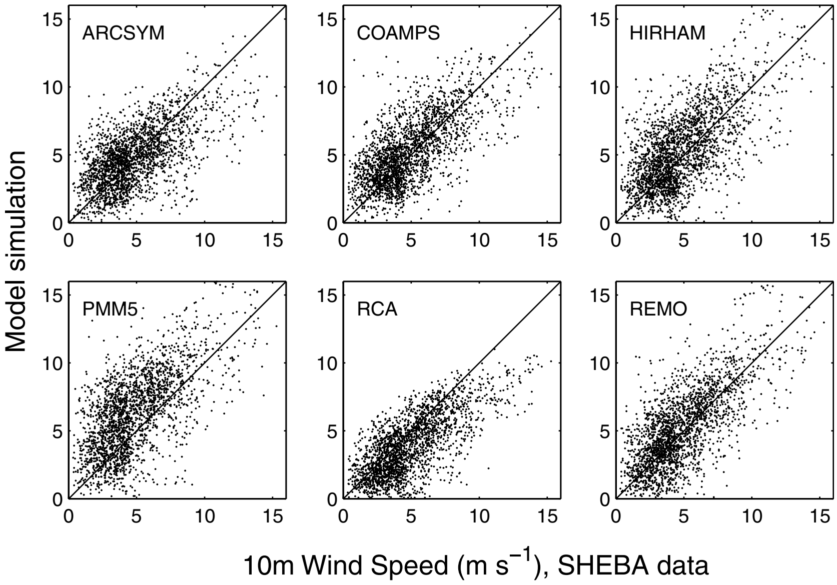

Figure 7. Scatter plots of the modelled (vertical axis) against the observed (horizontal axis) wind speed at 10 m for the six models, in m s^{-1}.

through the 1998 season (not shown), allows calculation of a relative measure of the error. Represented this way, RCA and Polar MM5 are closest to the observations, with a 4 and 1% error, respectively. The other models are low, with errors of −8, −15, −18% and −27% for ARCSyM, COAMPS$^{\text{TM}}$, HIRHAM and REMO, respectively.

An accurate upward flux (Figure 8b; Table III) requires both an accurate downward flux and an accurate surface albedo. The models with the lowest incoming shortwave radiation are also those with the lowest outgoing shortwave radiation. REMO is again markedly low, from mid-summer through the rest of Arctic summer. Polar MM5 and RCA are slightly high, while COAMPS$^{\text{TM}}$ is marginally closer to the measurements than for its downward flux. In an annually accumulated sense, the models are more scattered around the observed results than for the downward flux. REMO is still the lowest by far, with an error of −38% relative to the annually accumulated flux. COAMPS$^{\text{TM}}$ is now the best with a −4% error, while RCA, Polar MM5 and ARCSyM are too high, by 21%, 16% and 7%, respectively. Thus, in some models, the errors in upward and downward fluxes compensate each other, while in others there is a remaining net error. For both the upward and the downward fluxes, the modelled standard deviation is similar to the observed, while the RMS error is about half the standard deviations, except for HIRHAM and REMO, which have slightly higher RMSE in

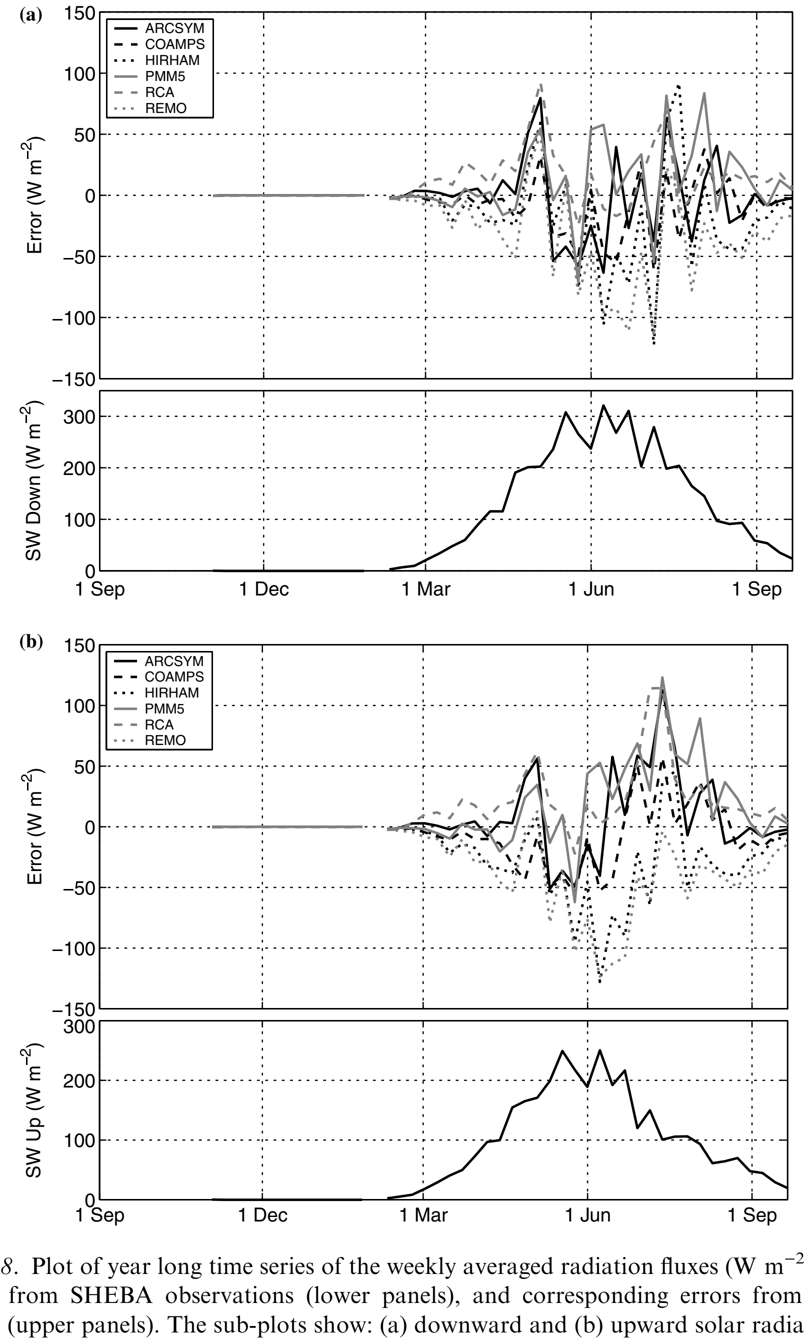

Figure 8. Plot of year long time series of the weekly averaged radiation fluxes (W m^{-2}) and of albedo from SHEBA observations (lower panels), and corresponding errors from the six models (upper panels). The sub-plots show: (a) downward and (b) upward solar radiation, (c) albedo, and (d) incoming longwave thermal radiation. Note that two albedo values are given in the sub-plot (c); the solid line is the point measurement from the ASFG site and the dashed is a line average. The legends are the same as in Figure 4: ARCSyM, COAMPS^{TM} and HIRHAM are black while Polar MM5, RCA and REMO are grey, solid, dashed and dotted, respectively.

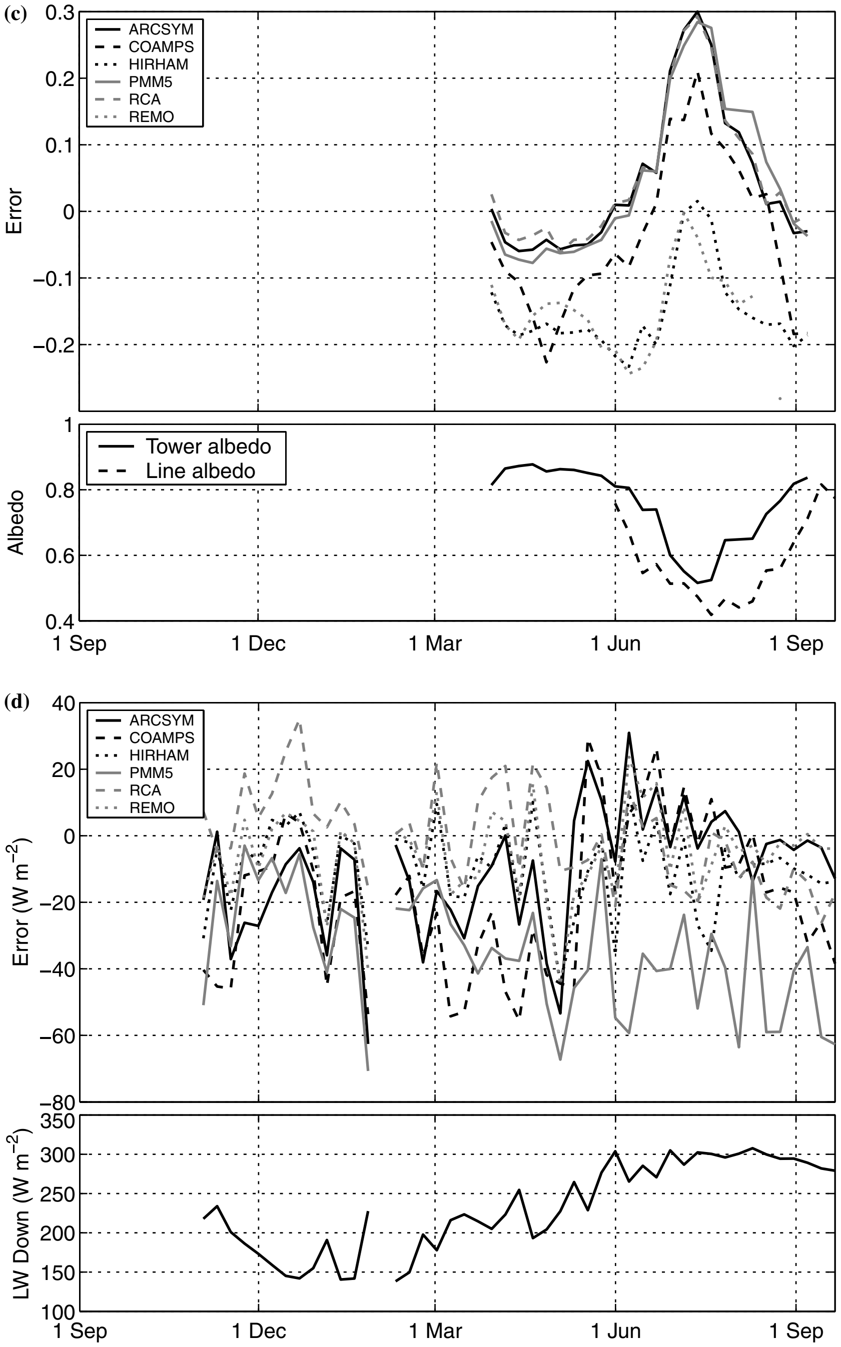

Figure 8. Continued.

relation to the standard deviation. Both the correlation coefficients and the IoA are relatively high, 0.8 and 0.9, respectively, or somewhat larger, with COAMPS^{TM} the highest and HIRHAM slightly lower.

The observed and calculated albedos are shown in Figure 8c. Model simulated albedos are calculated from the ratio of the modelled modelled upward to downward radiation at the surface and the errors are calculated using the ASFG albedo, which is a point measurement. The line-averaged albedo, which accounts more realistically for melt-pond distribution (Persson et al., 2002) making it more comparable to the area-averaged model values, is not available for the whole season. Models with a positive error are thus worse than indicated by the errors shown in these plots and *vice versa*, since the line-averaged albedo is lower. Inspecting the modelled albedo, all models fail in some respect. HIRHAM and REMO are both too low and capture the summer albedo only for a short while, but have a much too low albedo during spring and autumn. ARCSyM, RCA and Polar MM5 have a relatively constant albedo, in fact the Polar MM5 albedo is exactly constant, and are reasonable during spring and autumn, but have a much too high albedo during the summer. These models, with the exception of Polar MM5, have a slight seasonal variation, but with an amplitude that is only half or less of what is observed. COAMPS^{TM} has a very small seasonal variation and varies more with events of new snow. The end results give errors in net shortwave radiation roughly from -13 to 14 W m^{-2}.

The relative agreement between all models in upward longwave flux (not shown) simply reflects the prescribed surface temperature. It is somewhat too high during winter and somewhat low in summer in most models, consistent with the errors in surface temperature discussed earlier. From early March through early June, the results on a weekly basis are almost perfect in all models. The annual bias is negative and small, less than -10 W m^{-2}, except in COAMPS^{TM} and Polar MM5 (Table III). As for the incoming shortwave radiation, the incoming longwave radiation (Figure 8d) is also sensitive to the presence of low-level clouds, but their thickness should not be critical, since they become ‘black bodies’ even for rather shallow clouds. The annual cycle in incoming longwave radiation is modelled reasonably well by all models and the inter-model scatter is smaller than that for incoming shortwave radiation. Incoming longwave radiation is markedly lowest in Polar MM5, with the largest bias in summer, with errors of up to -50 W m^{-2}, and an annual mean bias about -35 W m^{-2}. Summer is the period when the inter-model spread among the remaining models is smallest and errors mostly around zero or somewhat less; the remaining models agree well with the measurements. During winter, the spread is larger and ARCSyM and sometimes COAMPS^{TM} are almost as low as Polar MM5. These, together with HIRHAM, also have the next largest negative annual bias, between about -35 and -10 W m^{-2}. RCA has the largest incoming longwave radiation during the winter, but has the smallest bias on an annual basis, close to zero. In an integrated sense through the SHEBA time period, all models are good to within 5%, and the error statistics are also good. RMS errors are typically smaller than the observed and modelled standard deviations, which are similar, and the correlation coefficients and the IoAs are larger than 0.8 and 0.9, respectively, although somewhat lower for Polar MM5. In terms of net radiation, the longwave errors are slightly larger than for shortwave radiation, ranging from −34 to near zero W m^{-2}.

Figure 9 shows a scatter plot of the simulated downwelling longwave radiation compared to the measured radiation for the individual models. There is a lower limit to how low this radiation can reasonably become during cloud free conditions at the very lowest winter temperatures. Similarly, there is also an upper limit during the summer with low-level air temperatures near 0 °C and overcast low-level cloud conditions. Several models show a preference for these two states. Polar MM5 and COAMPS^{TM} underestimate and REMO slightly overestimates the radiation at the lower limit, while there is a tendency for all models to underestimate the flux at the upper limit. The models also have distinctly different structure for the intermediate range. ARCSyM, for example, has a bimodal structure with very few cases with a simulated incoming longwave radiation around 200 W m^{-2}. We speculate that the cause lies in how fractional clouds are

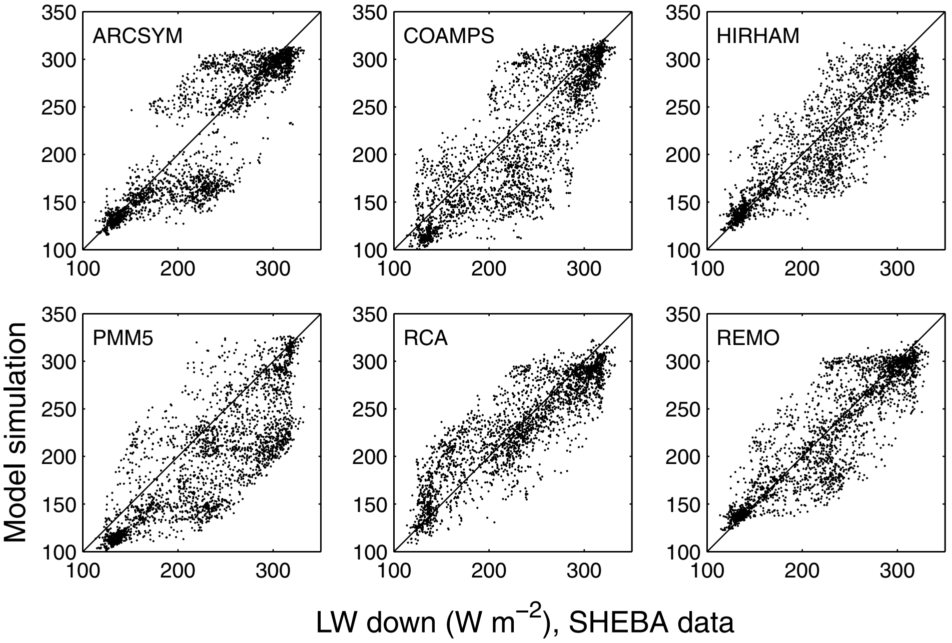

Figure 9. As Figure 7, but for the incoming longwave thermal radiation, in W m^{-2}.

treated in the different models. It also seems common among these models to have biases of different structure for simulated radiation at the maximum and minimum values. This error structure indicates that the models have problems simulating the clouds for different seasons. ARCSyM, for example, seems to overpredict low clouds in summer but underpredicts them in winter, while COAMPS^{TM} seem to somewhat underpredict and REMO overpredict summer low clouds, with better results for winter. RCA also underpredicts summer low clouds but overpredicts winter low clouds. Only HIRHAM seems to have a randomly scattered error.

3.4. Surface turbulent fluxes

Time series of the turbulent heat fluxes, both defined to be positive upward, are shown in Figure 10. Simulated as well as measured fluxes are relatively small throughout the whole year. The modelled weekly-averaged sensible heat flux varies between −30 and 20 W m^{-2}, while the corresponding latent heat flux varies between −5 and 20 W m^{-2} (Figure 10a and b). The observed values are smaller in magnitude and the observed latent heat flux is almost never negative. Overall, however, the magnitudes of the simulated fluxes are not inconsistent with the observations. The individual annual-averaged model biases are small, ranging from about −4 W m^{-2} (RCA) to slightly above 1 W m^{-2} (REMO) for sensible heat and from almost 0 W m^{-2} (Polar MM5) to about 6 W m^{-2} (ARCSyM) for latent heat. The striking feature is instead the very large variability both within and between the models. The standard deviations of the modelled heat fluxes are consistently larger than those observed, with the largest difference in ARCSyM and RCA, by a factor of 2-5 (Table IV). The correlation coefficients are mostly about 0.1 to 0.2 and 0.1 to 0.3 for sensible and latent heat, respectively. The highest correlations are found in COAMPS^{TM}, with 0.39 and 0.46 for sensible and latent heat flux, respectively. Although the IoA is generally somewhat higher in all of the models, it is only in COAMPS^{TM} that it is as large as 0.6.

At a first glance, one may be tempted to conclude that all this is less important since the fluxes are so small. However, the net heat flux at the surface is often of the same magnitude as these errors. Based on hourly SHEBA measurements, the net heat flux at the surface is in the range ± 20 W m^{-2} for about 45% of the time (excluding heat conduction through the snow and ice, not shown). Thus, even errors of a few tens of W m^{-2} may be significant in the Arctic. The fact remains that the fluxes from the individual models have very little resemblance to each other or to the measurements. Dethloff et al. (2001) also showed very different results, while experimenting with a model with different parameterisations, although no direct evaluation of the fluxes was performed.

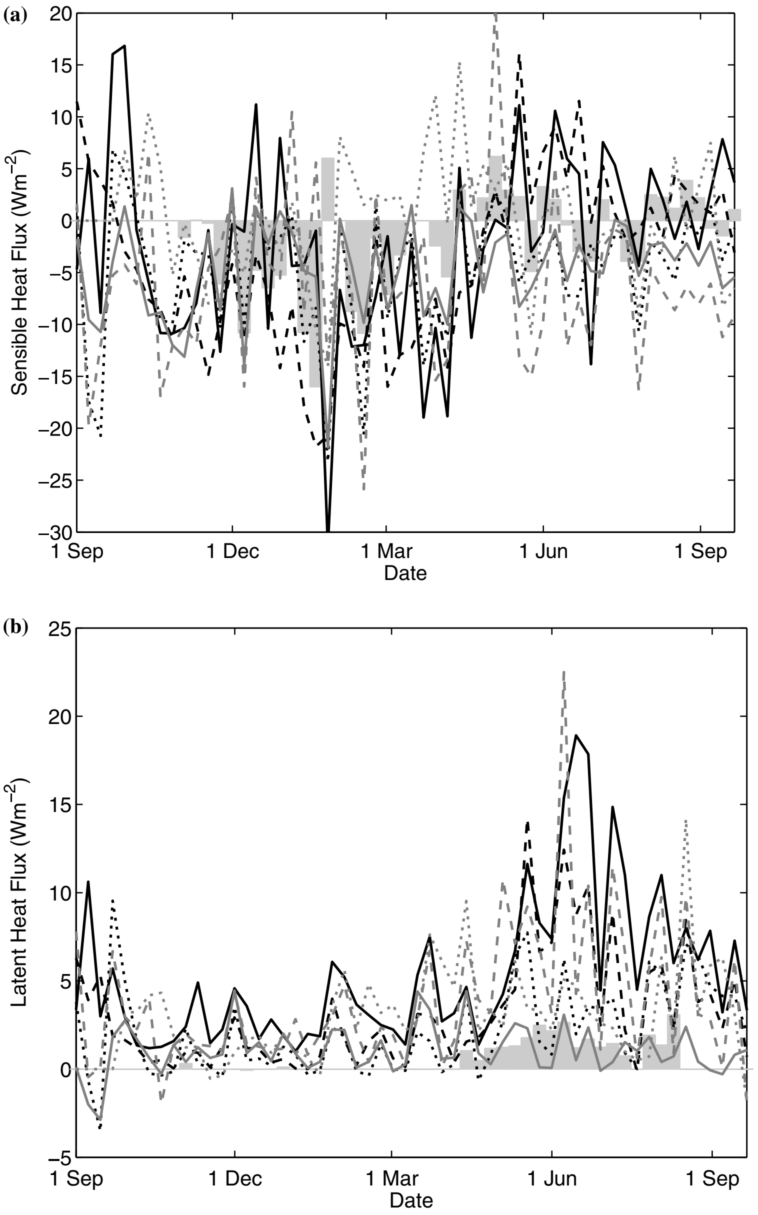

Figure 10. Plots of year long time series of weekly averaged (a) sensible and (b) latent turbulent heat flux (W m^{-2}) for all six models and the SHEBA observations (see the legend in Figure 4). The accumulated values of the sensible and latent heat fluxes are shown in (c) and (d), respectively. Note that in the accumulations, each 3-hourly value is considered representative for the whole corresponding 3-hour period, both from the models and from the observations. The legend is the same as in Figure 4: ARCSyM, COAMPS$^{\mathrm{TM}}$ and HIRHAM are black while Polar MM5, RCA and REMO are grey, solid, dashed and dotted, respectively. Grey vertical bars represent SHEBA.

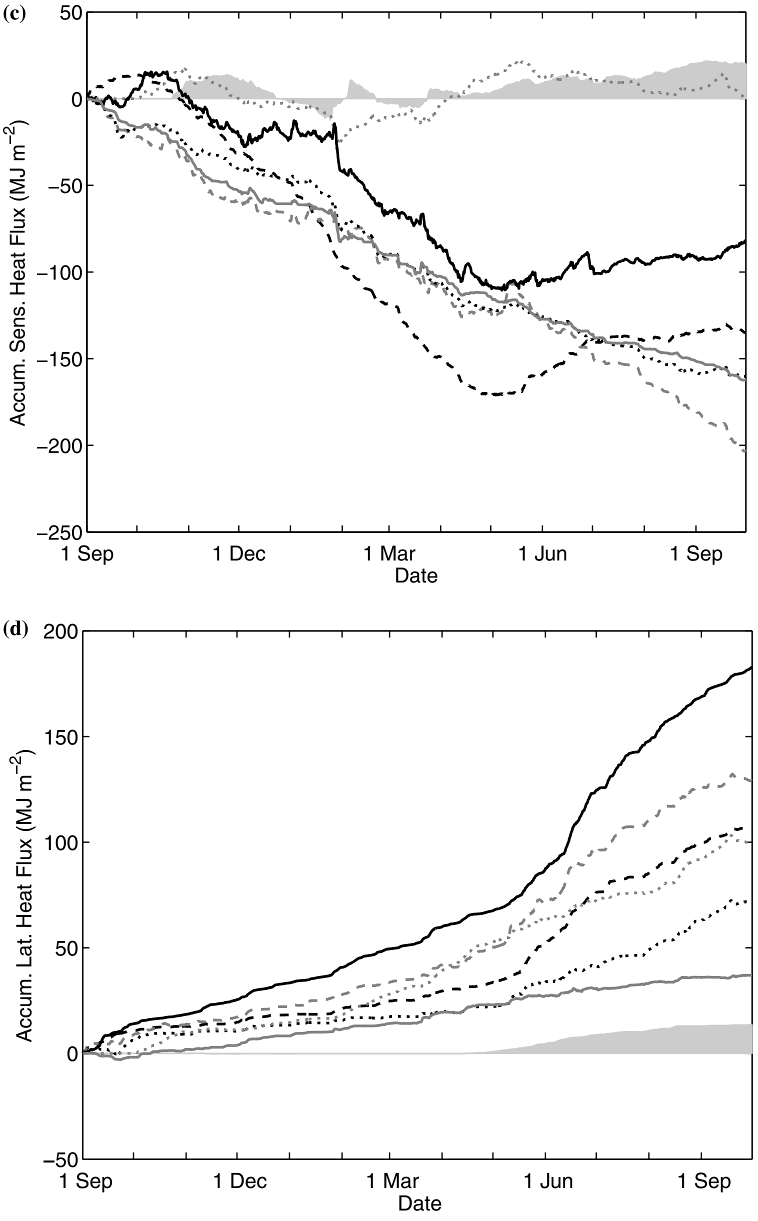

Figure 10. Continued.

Another way to analyse these fluxes is assess their accumulation over the year (Figure 10c and d), where it immediately becomes clear that the problem is severe, in particular if considered in the context of coupled modelling. Most of the models accumulate the negative sensible heat flux to an order of magnitude or more larger than the measurements. For some models, this is to some degree compensated for by the accumulated latent heat flux error, however, the inter-model spread in latent heat flux is even larger and the net balance thus varies between models. For example, REMO has the smallest accumulated sensible heat flux error, but still has a large accumulated latent heat flux error. In general, models with a large accumulated sensible heat flux do not necessarily have a large but opposite accumulated latent heat flux, although the net error is probably cancelled in the model ensemble average. Thus, a net balance for an individual model is not achieved.

The turbulent momentum flux (or the friction velocity, u_*) is crucial for the near-surface wind speed and also for the production of turbulence and therefore for all other turbulent fluxes. Additionally, an incorrect surface friction will bias the strength of synoptic-scale cyclones and anticyclones, by altering the cross-isobaric flow at the surface and thus the secondary circulation, giving rise to so-called 'spin-up' and 'spin-down' (e.g. Holton, 1992). The modelled friction velocity agrees much better with observations than the heat fluxes (Figure 11), though most models have a somewhat too high friction velocity. RCA has the largest positive bias in u_*, by about 0.1 m s^{-1}, consistent with its too low wind speed. On the other hand, Polar MM5 with the largest positive bias in wind speed still has the smallest bias in u$_*$. Polar MM5, however, has a bootstrap lower limit where u_*^2 is constrained to >0.005 m^2 s^{-2}, making the results difficult to interpret. ARCSyM and REMO have the second largest positive bias of slightly less than 0.1 m s^{-1}, however, without significant biases in the wind speed, while the remaining models are similar with a smaller positive bias. The modelled standard deviations are slightly larger than those observed but the RMS errors are of about the same magnitude. The correlation coefficients are between 0.60 (Polar MM5) and 0.68 (COAMPS$^{\mathrm{TM}}$) while the IoA's are slightly larger, 0.7 to 0.8.

Most models have complex systems of compensating errors and are often inadvertently tuned so that systematic errors cancel between different processes in a complex system that becomes very difficult to penetrate (e.g. Randall and Wielicki, 1997; Randall et al., 2003). Often the surface flux parameterisations were developed in weather forecast models and are consequently often inspired by overall model performance than by conformity to observations. Therefore, investigating functional behaviours between different variables in the models is often useful. Figure 12 shows u_* plotted against the 10-m wind speed for models and for measurements, inspired by the bulk-flux framework, where u_* should be proportional to wind speed and the slope is proportional to the drag coefficient (with a slight stability correction). This relationship is quite different in the different models. First, the scatter is much larger in some models than in others, for example ARCSyM has a much larger scatter than REMO and HIRHAM, with almost no scatter at all. With

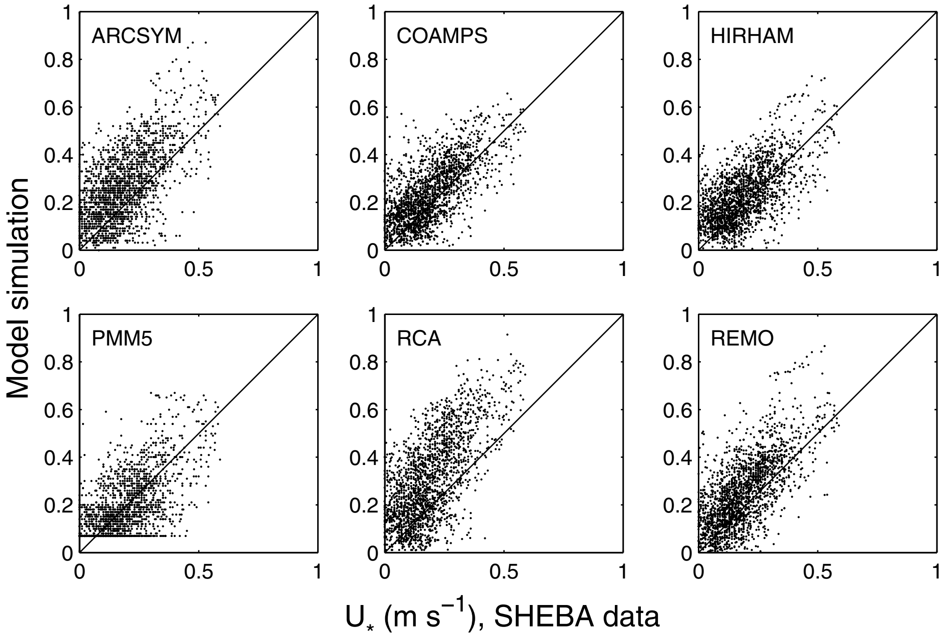

Figure 11. As Figure 7, but for the friction velocity, u_*, in m s^{-1}.

the possible exception of ARCSyM, the scatter is smaller in all of the models than in the observations. Second, the scatter is sometimes larger for low values of u_*, for example in COAMPS$^{\mathrm{TM}}$ and Polar MM5, while in other models the scatter is about the same for all values (e.g., ARCSyM). For Polar MM5 there are occasions with a quite high wind speed, about 10 m s^{-1}, and still a very small friction velocity being maintained at the lower threshold. There is no particular organisation to the scatter when u_* is analysed according to stability (not shown). Third, the value of u_* for a given wind speed varies in different models, implying that the drag coefficients used here are markedly different. At, for example, a wind speed of 10 m s^{-1}, u_* varies by a factor of two from about 0.4 m s^{-1} (HIRHAM and Polar MM5) to about 0.8 m s^{-1} (RCA). It is likely that the large stress in RCA for a given wind speed explains its low bias in wind speed, probably goes back to the use of different roughness lengths.

Staying within the bulk-flux formulation framework, the sensible heat flux should be proportional to both the wind speed and to the air-surface temperature difference. Similarly, latent heat flux should to a first order be proportional to wind speed and the air-surface moisture difference, where the surface moisture should here be at ice saturation with respect to the surface temperature. Plotting sensible heat flux divided by scalar wind speed against

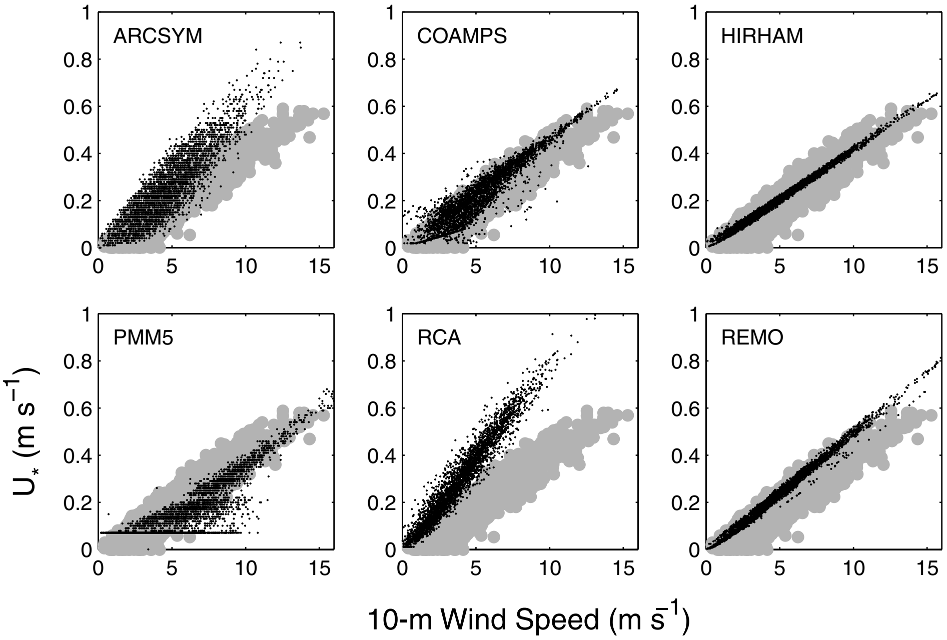

Figure 12. Scatter plots of the modelled friction velocity, u_*, (vertical axis) against the modelled wind speed at 10 m (horizontal axis), both in m s^{-1}, for the six models. The grey area represents the ensemble of the same relationship in the observations.

the temperature difference between 2 m and the surface (Figure 13), the slope should be determined by a heat transfer coefficient. In the observations there seems to be two different regimes. In one, the scaled flux has an almost linear dependence to the temperature difference, indicating a (nearly) single-valued transfer coefficient. All models capture this regime reasonably well, although the internal model scatter is different in different models. The slopes of the dependence are closest to the observations in ARCSyM and HIRHAM; COAMPS$^{\mathrm{TM}}$ and REMO have too small a slope while Polar MM5 and RCA have too large a slope. In all models, the scaled heat flux cover roughly the same range as in the measurements. This means that if the magnitude of the actual heat flux in Figure 10 is outside the largest values in the measurements, to a first approximation this has to be caused by an error in wind speed, rather than in the atmosphere–surface temperature difference or in the surface-flux parameterisation. The other regime is entirely confined to positive temperature differences and relates to the more stable conditions, indicating non-unique scaled heat fluxes for a given temperature difference. Further analysis shows that this can be interpreted as the heat-flux coefficient being a function of Richardson number. Only COAMPS$^{\mathrm{TM}}$ shows anything resembling the observed second regime. A closer look at the COAMPS$^{\mathrm{TM}}$

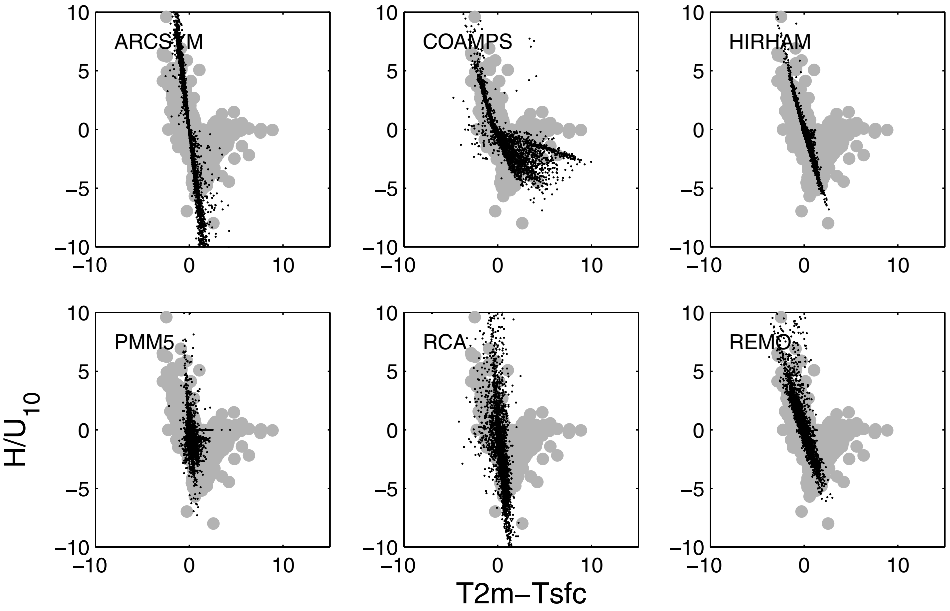

Figure 13. The sensible heat flux scaled with the 10-m wind speed (N m^{-2}) against the temperature difference between 2 m and the surface (K), for each model. Note that here we have used the apparent surface temperature, as it is used in each model. The grey area represents the ensemble of the same relationship in the observations.

results reveals that the transfer coefficient here has a dependence on the surface Richardson number for stable stratification that seems to be absent in all of the other models.

For the scaled latent heat flux (Figure 14), the differences between the models are larger and the results harder to interpret. The three models that seem to mimic the observed structure best are ARCSyM, COAMPS$^{\mathrm{TM}}$, and HIRHAM, although the scatter in ARCSyM is much larger due to scaled fluxes that are too large. RCA also has a large scatter with much too large values of the scaled flux. In Polar MM5, there is almost no dependence at all in the scaled flux to the moisture difference, while RCA seems to have an overly large dependence on this difference. The REMO results have the smallest moisture difference distribution of all of the models. Thus, in contrast to the sensible heat flux, the modelled scaled latent heat flux is frequently larger than the observations. The erroneously large values of the latent heat flux itself (Figure 10b) is thus not only dependent on wind speed errors.

4. Vertical Structure

Evaluating the vertical structure in model simulations is significantly more difficult, due to the lack of high-quality measurements, and the present

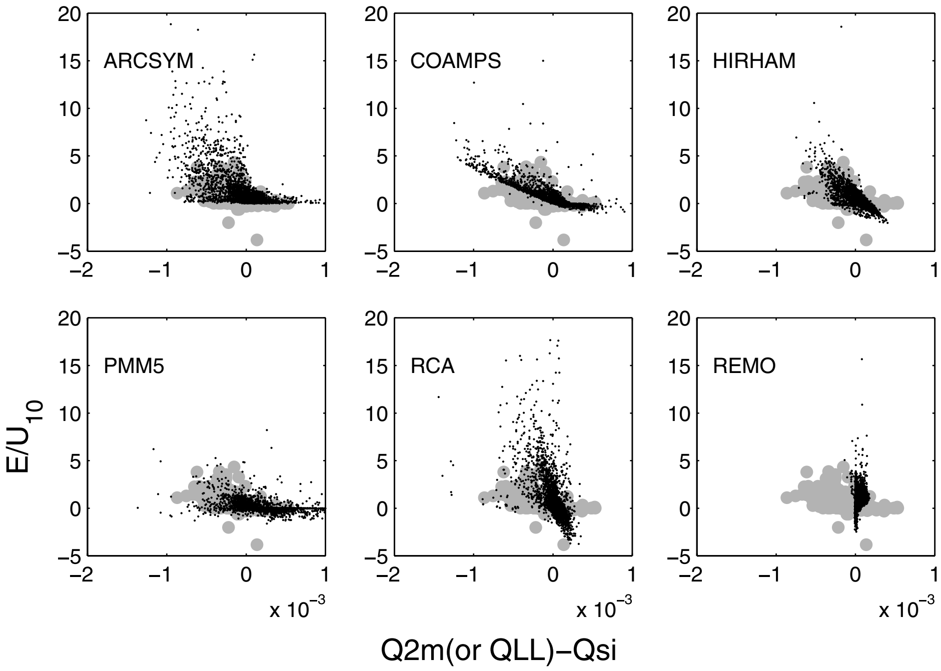

Figure 14. Same as Figure 13, but for the scaled latent heat flux (N m^{-2}) against the low-level specific humidity difference (kg kg^{-1}). Note that the latter is taken between 2 m and the surface for HIRHAM, RCA and REMO, while for the other models the lowest model level is used. The surface moisture is the saturation value with respect to ice at the surface temperature.

evaluation is based on the regular SHEBA soundings. During SHEBA other low-level soundings were also deployed, however, the regular soundings are the only available measurements of vertical temperature, moisture and windspeed profiles through the boundary layer with adequate data coverage for a systematic analysis. The top panel of Figure 15 shows a time-height cross-section of temperature from soundings for the whole SHEBA year. A striking feature is the frequent intrusions of warm (and also moist, not shown) air occurring at the top of the boundary layer throughout the year. They appear to be more frequent but shorter in duration in winter but more persistent in summer. The lower panel in Figure 15 shows the ensemble-averaged model bias (shaded) and the inter-model standard deviation (contours).

Two different ‘types’ of ensemble-average bias are identifiable. In one type, the bias is highly coherent in the vertical but has a relatively short time scale, see for example early November, early March, early April, and late July. This bias occurs close to the major warm-air intrusions and is also co-located with peaks in inter-model variability. It seems likely that this type of error is due to errors in timing. The warm-and-moist intrusions start and end with the

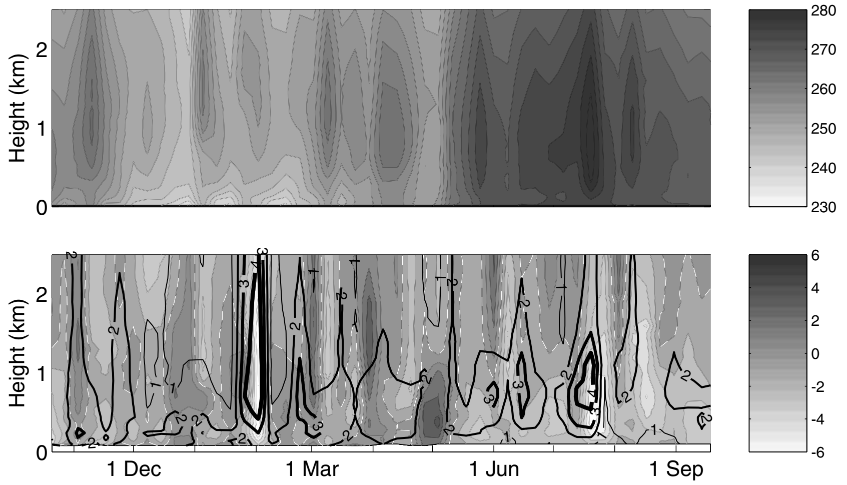

Figure 15. Time-height cross-sections of the observed temperature (K), from SHEBA soundings (top panel), and a combination of the model ensemble-average bias (grey shaded) and the inter-model standard deviation (contours) in the bottom panel.

arrival of frontal zones, with a pronounced vertical coherence. A slightly different arrival time of these fronts in the different models will induce a large inter-model variability. The ensemble-average arrival time is probably also slightly in error; this could be due to the small ensemble, but also to inadequate resolution or errors in the driving global model analysis. This type of error is less worrying from a climate modelling perspective. In the other type, the time scale is longer but the bias is more confined to near the surface, for example late April to early May and late June. The vertical structure of this error is shallower and mostly occurs below 1 km. These errors often tend to be negative (too cold) in summer and positive (too warm) in winter, and are not necessarily correlated with large inter-model variability. A systematic overestimation of low clouds could have this effect, but this is difficult to evaluate and cloud details are beyond the scope of this paper. Other errors also occur that probably require a case-by-case examination to understand.

Figure 16 shows seasonally averaged bias profiles of several variables from the six models. The temperature-bias profiles are shown in Figure 16a, where four of the models have a relatively consistent temperature error in the free troposphere through the year (N.B., errors above 4 km were not evaluated here): ARCSyM, RCA, Polar MM5 and HIRHAM, the latter two with a slightly larger interseasonal variability. In COAMPS$^{\mathrm{TM}}$ and REMO, the interseasonal variability is larger and in both, winter is significantly colder than the other seasons, spring and autumn are similar and in the middle of the range, and summer is too warm. In REMO, the biases seem constant above about 2 km, while errors continue to grow with height in

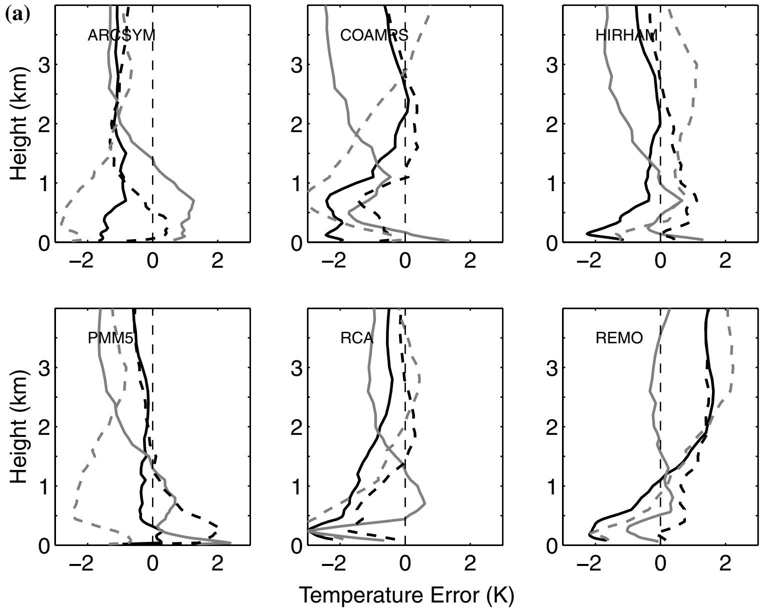

Figure 16. Seasonal averages of vertical bias profile between the surface and 4 km for (a) temperature (K), (b) specific humidity (g kg^{-1}) and (c) wind speed (m s^{-1}). Autumn and winter are represented by solid black and grey lines, respectively, and spring and summer by dashed black and grey lines, respectively.

COAMPS^{TM}, at least in summer. The bias ranges from zero to 2 °C in REMO and between −2 and 1 °C in COAMPS^{TM}. In all of the models, the bias in the lowest portion of the troposphere has a different structure compared to that higher up; this difference appears below roughly 1 km. Below 1 km, COAMPS^{TM}, HIRHAM, RCA and REMO are approximately consistent between the seasons. Their structure is also similar, with a cold-bias peak at some height below about 1 km and a relative warming closer to the surface, but the heights to the various peaks vary between the models. The low-level cold bias is at a greater height in COAMPS^{TM}, at about 1 km, while the peak of the cold bias is at only a few 100 m in some other models. In both ARCSyM and Polar MM5, winter and spring have a warm bias, while summer and autumn are biased colder compared to errors higher up in the troposphere. Of interest to note is that all of the models have significant peaks in average cloud water below 1 km, in particular in summer (not shown). In fact, in practically all of the models, the summer low-level peak in cloud water is almost a mirror image of the corresponding peak in negative temperature bias. Thus, again a closer investigation of the modelled low-level clouds is a logical next step. An evaluation of the modelled clouds is beyond the scope of this paper, but some comments are unavoidable. Modelled cloud

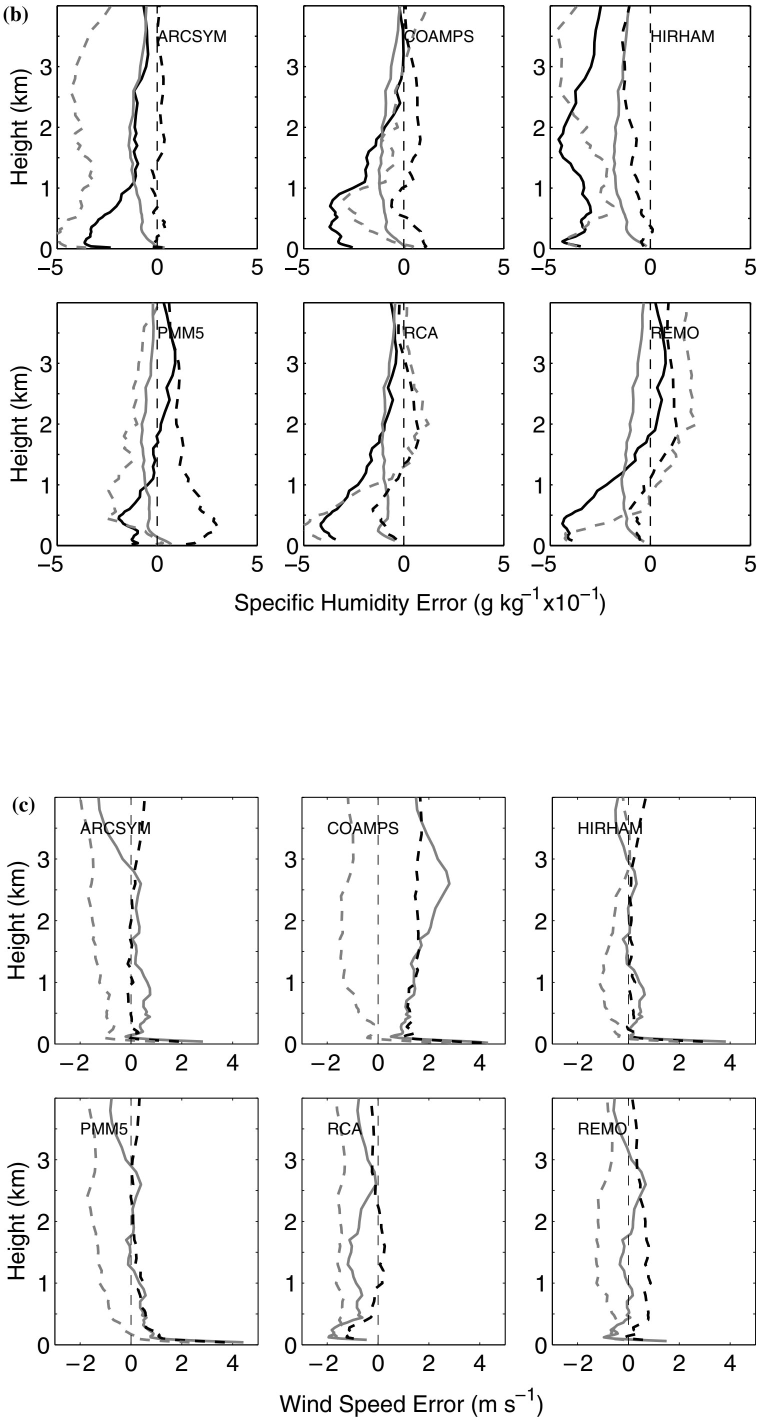

Figure 16. Continued.

water (not shown) is highly correlated with positive errors in water vapour in all the models, in particular in summer.

The humidity bias profiles (Figure 16b) have a somewhat different structure than that for temperature. Four models, COAMPS^{TM}, Polar MM5, RCA and REMO, have consistent near-zero biases above 2 or 3 km for all seasons, although REMO is somewhat more variable. HIRHAM is more variable through the seasons but is consistently too dry, by 0.1 to 0.4 g kg^{-1}. ARCSyM deviates in summer only, by about -0.4 g kg^{-1}, while the remaining seasons are closer to the measurements. The bias structure below 1 km is also different compared to temperature. Mostly, summer and autumn are drier than aloft relative to the observations, while in the other seasons the boundary-layer bias is roughly the same as aloft.

During autumn, an insufficient number of soundings with acceptable quality wind measurements preclude an evaluation of the wind-speed profiles for this season (Figure 16c). Moreover, strong winds (>30 m s^{-1}) are almost never observed, presumably due to instrument problems. In addition, the quality of the sounding of winds below 100 m is very questionable, thus the systematic rapid increase in bias closest to the surface in all models and all seasons can probably be ignored. The general impression is that all of the models are relatively close to the observations in winter and spring, except COAMPS^{TM}, which has wind speeds that are too high. All of the models have a negative bias in summer, by about 1 to 2 m s^{-1}. RCA in addition has a local negative bias in a shallow layer below 400 m, which reaffirms the previous conclusion that its negative bias in near-surface wind speed is related to an overly large momentum flux in the boundary layer. ARCSyM develops a winter negative bias above 3 km, while COAMPS^{TM} is biased high by 1 to 2 m s^{-1} in both seasons, but less closer to the surface.

5. Summary and discussion

Six state-of-the-art regional-scale models have been operated for the so-called SHEBA year and are evaluated against measurements. The study focuses on the lower troposphere: near-surface variables, surface turbulent and radiative fluxes, and the vertical structure of temperature, moisture and wind speed. Several different models are used to build an ensemble that will help generalise the results. In fact, this study shows that the bulk properties of all models are similar, with all models deviating substantially from the observations in some respect.

It is expected that near-surface temperature, and closely related variables, follow the measurements well, since the ice-surface temperature was prescribed. Errors in the 2-m air temperature, low-level moisture and mean sea-level pressure are also relatively small and show features expected from the known errors in the forcing fields. The errors in the 10-m wind speed are somewhat larger, on the order of less than ± 1 m s^{-1}. Here different models have different, but systematic, errors but all model winds follow the observed temporal development well. For these six models, the friction velocity itself is acceptable. While there are rather large differences between the models, all follow the same temporal trends and systematic errors are mostly consistent with other biases, e.g. in wind speed.

Surface heat flux suffers from two different types of errors. The radiation fluxes at the surface agree less well with the mesurements but, with a few exceptions, the models results are encouraging considering the complexity of the problem. Correct radiation fluxes at the surface require a correct cloud field, which is far from trival. A closer evaluation of the clouds is, however, outside of the scope of this paper. More serious is the fact that the net errors in the different models arise from different components of the radiation budget. The turbulent heat flux suffers from a different type of error and the situation is generally worse. While the annual mean biases are quite small, no model is similar to any other model and none of them shows much similarity to the observations. Thus, even if the bias is small, the correlation coefficients between modelled and observed turbelent heat fluxes are typically below 0.3.

Only one model is marginally less poor, that with the highest vertical resolution near the surface, illustrating the necessity of having the first grid point within the surface layer. Assuming the surface-layer depth is about 10% of the bondary-layer depth, (e.g., $0.2u_* f^{-1}$) the surface layer is shallower than 15 m (lowest grid point in COAMPS^{TM}) about 40% of the time, and almost never extends to 80 m (lowest grid point in RCA). While it could be argued that these errors are not important since the fluxes themselves are so small, it must be noted that the observes net heat flux at SHEBA is in the same range as these errors about half the time.

Functional dependencies for the turbulent friction show that at least two models have a significantly different drag coefficient than the others. For sensible heat flux the heat transfer coefficients are mostly similar and conforms to observations, except under very stable conditions. Only one model includes a reasonably realistic functional dependence for sensible heat flux under more stable conditions. The results for the latent heat flux are generally worse. Three of the models have acceptable dependencies compared with observations, while the remaining three have either too large or too small a sensitivity to changes in near-surface gradients.

Complicating the evaluation of vertical structure is the fact that model errors in the boundary layer are influenced by other errors emanating from the free troposphere. It is clear that the bias profileshave two different regimes, below and above 1 km. Errors below 1 km are generally larger and show larger scatter between season and between models, indicating problems with boundary-layer parameterisations and with boundary-layer clouds. Errors above 1 Km are smaller but also different between models, mostly in their seasonal dependence, pointing to other problems with the model parameterisations, since they were all forced with the same lateral boundary conditions.

The most significant conclusions from this study are:

1. *Dynamics*: Much of the resolved-scale meteorology is quite well captured by the models. There are biases in the near-surface wind speed that are consistent with biases in surface stress. Probably, the stress formulations have over time been adjusted to obtain a correct large-scale pressure field, by providing the necessary cross-isobaric mass flux in the boundary layer, rather than to provide an accurate surface stress.

2. *The energy balance*: There are errors in the surface heat fluxes that are often at least as large as the net heat flux itself. Some of the models have good skill in some of the radiation flux components. A problem is that none of the models have good, or consistent, skill in all components. The turbulent heat fluxes have very little similarity to the observations at all. While the long-team errors in turbulent heat flux tend to compensate somewhat, this is probably a result of tuning these together without adjusting the surface stress (see above), rather than attempting to obtain correct fluxes.

3. *Clouds*: The modelled representation of clouds are beyond the scope of this paper, but the results here make it inevitable that clouds are mentioned, especially in conclusion. Clouds play an important role for the surface energy balance and thus for the boundary layer, however, there is an apparent inconsistency here involving the clouds. The simulated downward shortwave radiation suggests the cloud amounts are reasonably accurate, and there is a positive correlation between cloud water and specific moisture bias in all models. There is also a structural resemblance between the cloud liquid-water profiles and the temperature-bias profiles in summer. This issue certainly deserves a closer study.

4. *Error cascade*: There are very good reasons to assume that some of the errors in the boundary layer have their roots elsewhere in the model. Most of the systematic errors are different in the lowest kilometre than aloft, but they seldom approach zero with altitude, despite applying the same lateral boundary conditions to all models.

In summary, these results lead us to the conclusion that there are uncertainties in current modelling of Arctic climate processes that must be reduced by improving important process descriptions in climate models. Although it appears likely that climate change will be a severe problem in the Arctic, it is unlikely that formulations in current GCMs are in general much better than in the state-of-the-art mesoscale models evaluated here. Therefore, it would appear prudent to consider scenario results of future Arctic climate from GCMs, in particular ice-melt scenarios, with considerable caution.

Acknowledgements

MT, MZ and GS were supported by the Swedish Research Council under contract 2002-5610 and by the Swedish climate-modelling program SWECLIM. They are also grateful to the COAMPS-team at the Naval Research Laboratory in Monterey, CA, U.S.A. COAMPS$^{\text{TM}}$ was run at the National Supercomputer Centre (NSC) in Linköping, Sweden, JJC acknowledges support from National Science Foundation grant OPP-008429. KD, AR, CJ and KW were supported by the European Union Framework 5 project GLIMPSE (EVK2-CT-2002-00164). KD and AR are also grateful to Jens Christenssen at the Danish Meteorological Institute. SP and TS thanks Daniela Jacob, for providing the REMO model and the opportunity to participate in ARCMIP, and Ralf Podzun, for technical support. CJ and KW are grateful for the assistance by Anders Ullerstig and Ulrika Willén. All authors acknowledge the invaluable work invested in setting up the ARCMIP program, and in the generation of the boundary and forcing fields for this experiment. In particular, we are grateful to Judy Curry, Amanda Lynch, Elizabeth Cassano and Jeff Key. The authors are also grateful to the SHEBA Atmospheric Flux Group, mainly Ola Persson and Chris Fairall for access to observations, and to Chris Bretherton and Stefan de Roode for the compilation of the SHEBA soundings.

References

Andreas, E. L., Guest, P. S., Persson, P. O. G., Fairall, C. W., Horst, T. W., Moritz, R. E., and Semmer, S. R.: 2002, 'Near-surface Water Vapor Over Polar Sea Ice is Always Near Ice Saturation: *J. Geophys. Res.* **107**(C10), 10.1029/2000JC000411, 2002.

Battisti, C. M., Bitz, C. M. and Moritz, R. M.: 1997, 'Do General Circulation Models Underestimate the Natural Variability in the Arctic Climate?', *J. Climate*, **10**, 1909–1920.

Beesley, J. A., Bretherton, C. S., Jacob, C., Andreas, E. I., Intrieri, J. M. and Uttal, T. A.: 2000, 'A Comparison of Cloud and Boundary Layer Variables in the ECMWF Forecast Model with Observations at the Surface Heat and Energy of the Arctic (SHEBA) ice camp', *J. Geophys. Res.,* **105**(D10), 12337–12349.

Brinkop, S., and Roeckner, E.: 1995, 'Sensitivity of a General Circulation Model to Parameterizations of Cloud–Turbulence Interactions in the Atmospheric Boundary Layer', *Tellus*, **47A**, 197–220.

Bromwich, D. H., Cassano, J. J. Klein, T., Heinemann, G., Hines, K. M., Steffen, K., and Box, J. E.: 2001, 'Mesoscale Modeling of Katabatic Winds over Greenland with the Polar MM5', *Mon. Wea. Rev.* **129**, 2290–2309.

Christensen, J. H., Christensen, O. B., Lopez, P., Van Meijgaard, E., and Botzet, A.: 1996, *The HRIHAM4 regional atmospheric model.* DNMI Sci. Rep. 96-4, Danish Meteorological Institute, Copenhagen, 51 pp.

Christensen, J. H. and Kuhry, P.: 2000, 'High-resolution Regional Climate Model Validation and Permafrost Simulation for the East European Russian Arctic', *J. Geophys. Res.* **105**(D24), 29647–29658.

Cassano, J. J., Box, J. E., Bromwich, D. H., Li, L., and Steffen, K.: 2001, Evaluation of Polar MM5 Simulations of Greenland's Atmospheric Circulation', *J. Geophys. Res.* **106**, 33867–33,890.

Curry, J. A.: 1986, 'Interactions among Turbulence, Radiation and Microphysics in Arctic Stratus Clouds', *J. Atmos. Sci.* **43**, 90–106.

Curry, J. A., Hobbs, P. V., King, M. D., Randall, D. A., Minnis, P. Isaac, G. A., Pinto, J. O., Uttal, T., Bucholtz, A., Cripe, D. G., Gerber, H., Fairall, C. W., Garrett, T. J., Hudson, J., Intrieri, J. M., Jakob, C., Jensen, T., Lawson, P., Marcotte, D., Nguyen, L., Pilewskie, P., Rangno, A., Rogers, D. C., Strawbridge, K. B., Valero, F. P. J., Williams A. G., and Wylie, D.: 2000, 'FIRE Arctic Clouds Experiment', *Bull. Amer. Meteorol. Soc.* **81**, 5–29.

Curry J. A. and Lynch, A. H.: 2002, 'Comparing Arctic Regional Climate Models', *EOS Trans.*, **83**, 87.

Cuxard, J., Bougeault, P., and Redelberger, J. L.: 2000, 'A Turbulence Scheme Allowing for Mesoscale and Large-eddy Simulations', *Quart. J. Roy. Meteorol. Soc.* **126**, 1–30.

Dethloff, K. C., Abegg, C., Rinke, A., Hebestadt, I., and Romanov, V.: 2001, 'Sensitivity of Arctic Climate Simulations to Different Boundary-layer Parameterizations in a Regional Climate Model', *Tellus* **53**A, 1–26.

Hanna, S. R.: 1994, 'Mesoscale Meteorological Model Evaluation Techniques with Emphasis on Needs of Air Quality Models', in R. A. Pielke, and R. P. Pearce, (eds), *Mesoscale modeling of the Atmosphere*, Meteorological Monographs, American Meteorological Society, Boston, USA, 47–62.

Hodur, R. M.: 1997, 'The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS)', *Mon. Wea. Rev.* **125**, 1414–1430.

Holton, J. R.: 1992, *An Introduction to Dynamic Meteorology*, Academic Press, San Diego, U.S.A, 511 pp.

Intrieri, J. M., Fairall, C. W., Shupe, M. D., Persson, P. O. G., Andreas, E. L., Guest, P. S., and Moritz, R. E.: 2002, 'An annual cycle of Arctic surface cloud forcing at SHEBA, *J. Geophys. Res.* **107**(C10), 8039, doi:10.1029/2000JC000439, 2002.

IPCC: 2001, *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change* [Houghton, J.T., Ding, Y., Griggs, D.J. Noguer, M., van der Linden, P.J., Dai, X., Maskell, K., and Johnson C.A. (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, U.S.A, 881 pp.

Jacob, D.: 2001, 'A Note to the Simulation of Annual and Interannual Variability of the Water Budget over the Baltic Sea drainage basin', *Meteorol. Atmos. Phys.* **77**, 61–73.

Jones, C. G., Wyser, K., Ullerstig, A., and Willén, U.: 2004, 'The Rossby Center Regional Atmospheric Climate Model. Part II: Application to the Arctic', *Ambio*, *in press.*

Key, J.: 2002, '*The Cloud and Surface Parameter Retrieval (CASPR) System for Polar AVHRR*', Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin, Madison, 59 pp.

Leck, C., Tjernström, M., Bigg, K., Matrai, P., and Swetlicki, E.: 2004, 'Microbes, Clouds and Climate: Can Marine Microorganisms Influence the Melting of the Arctic pack ice?', *Eos Trans.* **85**, 25–36.

Liu, J., Curry, J., Rossow, W., Key, J., and Wang, X.: 2005, 'Comparison of Surface Radiative Flux Data Sets over the Arctic Ocean', *J. Geophys. Res.*, **110**, C02015.

Louis, J. F.: 1979, 'A Parametric Model of Vertical Eddy Fluxes in the Atmosphere', *Boundary-Layer Meteorol.* **17**, 187–202.

Lynch, A. H., Chapman, W. L., Walsh, J. E., Weller, G.: 1995, Development of a Regional Climate Model of the Western Arctic. *J. Climate*, **8** 1555–1570.

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., Stouffer, R. J.: 2000. The Coupled Model Intercomparison Project (CMIP). *Bull. Amer. Meteorol. Soc.*, **81**, 313–318.

Mahrt, L.: 1998, 'Stratified Atmospheric Boundary Layers and Breakdown of Models', *Theoret. Comput. Fluid Dynamics* **11**, 263–279.

Mellor, G. and Yamada, T.: 1974, 'A Hierarchy of Turbulence Closure Models for Planetary Boundary Layers', *J. Atmos. Sci.* **31**, 1791–1806.

Mellor, G. L. and Yamada, T.: 1982, 'Development of a Closure Model of Geophysical Flows', *Rev. Geophys. Space Physics* **20**, 851–875.

Overland, J. E., McNutt, S. L., Groves, J., Salo, S., Andreas, E. L., and Persson, P. O. G.: 2000, 'Regional Sensible and Radiative Heat Flux Estimates for the Winter Arctic during the Surface Heat Budget of the Arctic Ocean (SHEBA) experiment', *J. Geophys. Res.* **105**(C6), 14093–14102.

Perovich, D. K., Andreas, E. L., Curry, J. A., Eiken, H., Fairall, C. W., Grenfell, T. C., Guest, P. S., Intrieri, J., Kadko, D., Lindsay, R. W., McPhee, M. G., Morison, J., Moritz, R. E., Paulson, C. A., Pegau, W. S., Persson, P. O. G., Pinkel, R., Richter-Menge, J. A., Stanton, T., Stern, H., Sturm, M., Tucker and W. B., Uttal, T.: 1999, 'Year on Ice Gives Climate Insights', *Eos Trans.* **80**, 483–486.

Persson, P. O. G., Uttal, T., Intrieri, J. M., Fairall, C. W., Andreas, E. L., and Guest, P. S.: 1999, 'Observations of Large Thermal Transitions during the Arctic Night from a Suite of sensors at SHEBA. Preprints, 3rd Symp. on Integrated Observing Systems., Jan. 10–15, 1999, Dallas, TX, 171–174.

Persson, P. Ola G., Fairall, C. W., Andreas, E. L., Guest, P. S., and Perovich, D. K.: 2002, 'Measurements near the Atmospheric Surface Flux Group tower at SHEBA: Near-surface Conditions and Surface Energy Budget', *J. Geophys. Res.* **107**(C10), 8045, doi:10.1029/2000JC000705, 2002.

Pinto, O. J., Curry, J. A., and Lynch, A. H.: 1999, 'Modeling Clouds and Radiation for the November 1997 Period of SHEBA Using a Column Climate Model', *J. Geophys. Res.* **104**, 6661–6678.

Randall, D. A., and Wielicki, B. A.: 1997, 'Measurements, Models, and Hypotheses in the Atmospheric Sciences', *Bull. Amer. Meteorol. Soc.* **78**, 399–399.

Randall, D. A., Krueger, S., Bretherton, C., Curry, J., Duynkerke, P., Moncrieff, M., Ryan, B., Starr, D., Miller, M., Rossow, W., Tselioudis, G., and Wielicki, B.: 2003, 'Confronting Models with Data: The GEWEX Cloud Systems Study', *Bull. Amer. Meteorol. Soc.* **84**, 455–469.

Räisänen, J.: 2001, 'CO_2-induced Climate Change in the Arctic area in the CMIP2 Experiments', *SWECLIM Newsletter* **11**, 23–28.

Rinke, A., Dethloff, K., and Christensen, J. H.: 1999, 'Arctic Winter Climate and its Interannual Variation Simulated by a Regional Model', *J. Geophys. Res.* **104**(D16), 19027–19038.

Rinke, A., Lynch, A. H., and Dethloff, K.: 2000, 'Intercomparison of Arctic Regional Climate Simulations: Case Studies of January and June 1990', *J. Geophys. Res.* **105**, 29669–29683.

Rinke, A., Gerdes, R., Dethloff, K., Kandlbilder, T., Karcher, M., Frickenhaus, S., Koeberle, C., and Hiller, W.: 2003, 'A Case Study of the Anomalous Arctic Sea Ice Conditions during

1990: Insights from Coupled and Uncoupled Climate Model Simulations', *J. Geophys. Res.* **108(**D9), 4275, doi:10.1029/2002JD003146, 2003.

Tjernström, M.: 2005, 'The Summer Arctic Boundary Layer during the Arctic Ocean Experiment 2001 (AOE-2001)', *Boundary-Layer Meteorol.* **117**, 5–36.

Tjernström, M., Leck, C., Persson, P. O. G., Jensen, M. L., Oncley, S. P., and Targino, A.: 2004, 'The Summertime Arctic Atmosphere: Meteorological Measurements during the Arctic Ocean Experiment 2001 (AOE-2001)', *Bull. Amer. Meteorol. Soc.* **85**, 1305–1321.

Vihma, T., Hartman, J., and Lûpkes, C.: 2003, 'A Case Study of an On-Ice Flow Over the Arctic Marginal Sea Ice Zone', *Boundary-Layer Meteorol.* **107**, 189–217.

Walsh, J. E., Kattsov, W. M., Chapman, W. L., Govorkova, V., and Pavlova, T.: 2002, 'Comparison of Arctic Climate by Uncoupled and Coupled Global Models', *J. Climate*, **15**, 1429–1446.

Zilitinkevich, S. S.: 2002, 'Third-order Transport due to Internal Waves and non-local Turbulence in the Stably Stratified Surface Layer', *Quart. J. Roy. Meteorol. Soc.* **128**, 913–926.