JAMES Journal of Advances in Modeling Earth Systems

AGU ADVANCING EARTH AND SPACE SCIENCE

RESEARCH ARTICLE
10.1029/2020MS002144

Special Section:
The Arctic: An AGU Joint Special Collection

Key Points:

- A set of "coupling-strength" diagnostics is presented for use in the evaluation and development of Earth System Models
- These diagnostics are used to link an Arctic-wide warm bias in the ECMWF operational model to the use of a single-layer snow model
- They are used to demonstrate that a multilayer snow model improves this by reducing the coupling strength to the land surface

Supporting Information:

- Supporting Information S1

Correspondence to:
J. J. Day,
jonathan.day@ecmwf.int

Citation:
Day, J. J., Arduini, G., Sandu, I., Magnusson, L., Beljaars, A., Balsamo, G., et al. (2020). Measuring the impact of a new snow model using surface energy budget process relationships. *Journal of Advances in Modeling Earth Systems*, *12*, e2020MS002144. https://doi.org/10.1029/2020MS002144

Received 17 APR 2020
Accepted 5 SEP 2020
Accepted article online 10 SEP 2020

Measuring the Impact of a New Snow Model Using Surface Energy Budget Process Relationships

Jonathan J. Day[1], Gabriele Arduini[1], Irina Sandu[1], Linus Magnusson[1], Anton Beljaars[1], Gianpaolo Balsamo[1], Mark Rodwell[1], and David Richardson[1]

[1]European Centre for Medium-Range Weather Forecasts, Reading, UK

Abstract Energy exchange at the snow-atmosphere interface in winter is important for the evolution of temperature at the surface and within the snow, preconditioning the snowpack for melt during spring. This study illustrates a set of diagnostic tools that are useful for evaluating the energy exchange at the Earth's surface in an Earth System Model, from a process-based perspective, using in situ observations. In particular, a new way to measure model improvement using the response of the surface temperature and other surface energy budget (SEB) terms to radiative forcing is presented. These process-oriented diagnostics also provide a measure of the coupling strength between the incoming radiation and the various terms in the SEB, which can be used to ensure that improvements in predictions of user-relevant properties, such as 2 m temperature, are happening for the right reasons. Correctly capturing such process relationships is a necessary step toward achieving more skilful weather forecasts and climate projections. These diagnostic techniques are applied to assess the impact of a new multi-layer snow scheme in the European Centre for Medium-Range Weather Forecasts'-Integrated Forecast System at two high-Arctic sites (Summit, Greenland and Sodankylä, Finland). A previous study showed that it will enhance 2 m temperature forecast skill across the Northern Hemisphere in boreal winter compared to forecasts with the single layer model, reducing a warm bias. In this study we use the diagnostics to show that the bias is improved for the right reasons.

Plain Language Summary Predicting air temperature near the surface on time scales from hours to decades ahead is of high importance to a wide range of end users. However, it is also extremely difficult to get right due to the large number of processes involved. Air temperature near the surface is affected by a large number of atmospheric processes such as turbulent mixing, radiation, cloud microphysics as well as land surface processes. As a result, systematic errors often have multiple causes and are hard to diagnose. Similarly, it can be hard to know whether improvements between forecast model versions occurred for the right reasons. This study presents a set of diagnostic tools that are useful for addressing this need. They are applied to assess the impact of a new snow model on experimental forecasts with the European Centre for Medium-Range Weather Forecasts' weather prediction system. They show that it improves forecasts of air temperature near the surface for the right reasons.

1. Introduction

Weather and climate models suffer from systematic errors in surface temperature and related heat fluxes (Zadra et al., 2018). This often leads to difficulties in predicting basic properties such as 2 m temperature, at time scales from minutes to decades, as highlighted by a recent survey of modeling centers conducted by the World Meteorological Organization's Working Group on Numerical Experimentation (WGNE, 2019). 2 m temperature (T_{2m}) forecast errors are particularly large when the boundary layer is stably stratified (e.g., Atlaskin & Vihma, 2012; Sandu et al., 2013), subsequently T_{2m} skill in polar regions is relatively low, in part, due to the prevalence of such conditions (Bauer et al., 2016; Jung et al., 2016). Moreover, even the most everyday of phenomena, the diurnal cycle of temperature in midlatitudes, has been hard to simulate, in part due to the sheer number of interacting processes (e.g., Lindvall & Svensson, 2015; Svensson et al., 2011).

The evolution of temperature in the atmospheric boundary layer is primarily influenced by atmospheric processes such as turbulent mixing, radiation, and clouds. However, coupling to the land surface also plays an important role, particularly during stable conditions, when turbulent exchange with the atmosphere is small

(Holtslag et al., 2013; Sterk et al., 2013). Therefore, because of the number of processes involved, systematic errors in forecasts of near-surface temperature at a given location, may have numerous causes (Haiden et al., 2018; Schmederer et al., 2019). Further, since errors in the representation of the various processes can compensate each other, T_{2m} skill may not necessarily be achieved for the right reasons. For example, a positive bias in incoming radiation could be compensated by excessive turbulent heat fluxes, resulting in the correct temperature.

The development of process-oriented diagnostics (PODs) for land-atmosphere coupling has tended to focus on the link between soil moisture and precipitation (Santanello et al., 2018), particularly in midlatitude continental regions, where the feedback between these two parameters are particularly strong and important for predictability (e.g., Koster et al., 2004, 2006). Numerical experimentation has been the dominant paradigm to identify sources of temperature error in stable boundary layers and in the diurnal cycle (e.g., Cuxart et al., 2006; Holtslag et al., 2013). However, more diagnostic-focused studies do exist. For example, the Clouds Above the United States and Errors at the Surface (CAUSES) project took a diagnostic approach to understanding the causes of error in summertime temperature over the U.S. Great Plains (e.g., Ma et al., 2018).

In this study we present a set of land-atmosphere coupling PODs designed to assess the response of surface temperature to radiative forcing in an Earth System Model. Errors in this response, broadly speaking, can be due to errors in the strength of coupling with the underlying medium (i.e., soil or snow) or to errors in the strength of coupling to the atmosphere (i.e., too much or too little diffusion). Both of these factors can have an impact on near-surface temperature forecast error (see Viterbo et al., 1999). The diagnostics presented here provide a way to quantify the strength of this coupling and compare this with observations.

The PODs presented in this study, which follow the ideas of Miller et al. (2018), are based on the idea that the surface energy budget

$$SW_{net} + LW\downarrow = -(SHF + LHF + GHF - LW\uparrow) \tag{1}$$

can be split into "driving terms": net shortwave radiation (SW_{net}) and incoming longwave radiation ($LW\downarrow$), and"'response terms": outgoing longwave radiation, $LW\uparrow$, and sensible, latent and ground heat fluxes (*SHF, LHF*, and *GHF*, all defined as positive when directed toward the surface). What distinguishes the driving terms from the response terms is that they are not directly dependent on the thermal properties of the surface. Miller et al. used the regression parameters between the driving term and the various response terms as a set of diagnostics which can be compared with observations and used to understand the causes of surface temperature error. They applied this technique to output from a climate model, a seasonal forecasting system and the ERA-Interim reanalysis (Dee et al., 2011), to diagnose the causes of low sensitivity of the surface temperature to variations in radiative forcing at the Greenland Summit Station which is a feature of all three datasets. In this study we explore how these techniques could be used to aid the model development process, using the example of a new multilayer snow model in the ECMWF forecast system.

Currently, most operational numerical weather prediction (NWP) models use only a single layer snow scheme (Essery, 2010) and as a result variations in snow temperature with depth cannot be captured. The importance of this vertical structure is illustrated by Figure 1, which shows the transition from warm-cloudy conditions to cold-cloud-free conditions at Sodankylä, Finland, in January 2014. During this period the cooling of the snowpack is largest and most rapid near the surface and the size and speed of snow temperature response reduces with increasing depth within the snowpack, with the snow closest to the soil hardly changing temperature due to the insulating effect of the snow above. From this it should be evident that with a single-layer snow model, it is impossible to simultaneously achieve both a realistic change in snow pack mean temperature and snow surface temperature, for a given change in radiative forcing. Indeed, the large thermal inertia associated with having to warm or cool the entire snowpack in the single-layer snow model used operationally in the European Centre for Medium-Range Weather Forecasts' (ECMWF) Integrated Forecast System (IFS) is thought to be a major cause of near-surface temperature errors in snow covered regions (e.g., Scandinavia, Haiden et al., 2018). This has led some climate models to introduce multilayer snow schemes, improving biases in the Northern Hemisphere (Walters et al., 2019). Similarly, it is expected that the inclusion of a multilayer snow scheme in the ECMWF-IFS will result in a more responsive surface

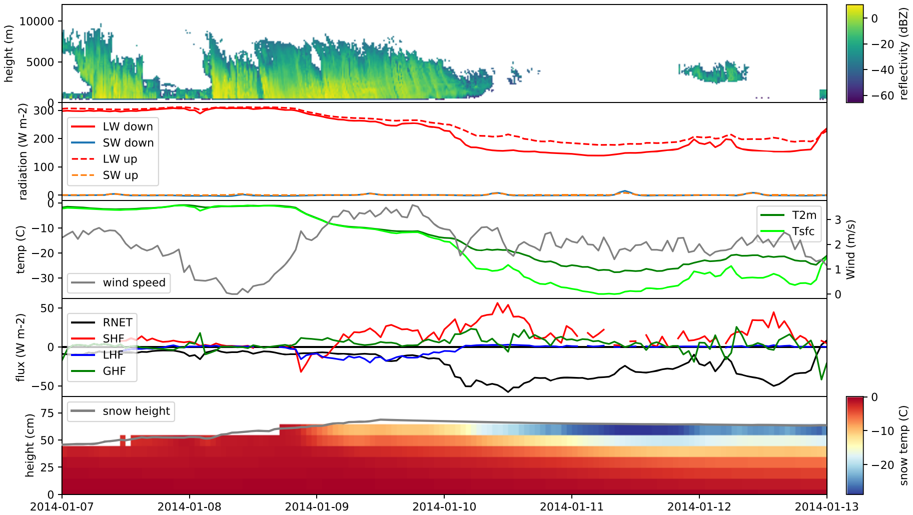

Figure 1. Observed meteogram for a case study at Sodankylä, Finland, in January 2014. It shows (from top-to-bottom) cloud radar reflectivity (from CloudNet; Illingworth et al. (2007)); radiation terms; wind speed, surface, and 2 m temperature; energy balance terms: Total net radiation (RNET), sensible (SHF), latent (LHF) and ground (GHF: Atmosphere snow) heat flux (with the sign convention that terms are positive when directed at the surface); and snow temperature at various heights (above the soil-snow interface).

temperature, especially for deep snowpacks. Directly representing a thin top layer, with a lower thermal inertia, will allow T_{sfc} to vary more in response to variations in radiative forcing than with the single-layer scheme. It will also influence the turbulent fluxes through their dependence on T_{sfc}.

A multilayer snow scheme was recently introduced in an experimental version of the ECMWF IFS (Arduini et al., 2019). They found that coupling to the new snow model reduced the bias in both 2 m temperature and snow depth overall, when compared to the conventional (SYNOP) observing network. However, there is a limit to what such evaluation can tell us about the processes responsible for those improvements, due to the limited set of parameters recorded at SYNOP sites. So called supersites, such as Sodankylä, Finland, and Summit, Greenland, on the other hand, collect a much wider set of observations, which can be used to evaluate model changes from a process-oriented perspective. In this study the PODs described above will be used to evaluate whether the improvements in 2 m temperature skill seen across the Arctic region in Arduini et al. (2019) are occurring for the right reasons and whether they are improving the overall behavior of the land surface-atmosphere interaction at those locations.

Although the analysis focuses on the impact of a new snow model in the Arctic during winter, we argue that the suite of PODs presented in this paper could be applied to any site with appropriate instrumentation for any season. They allow the impact of any model change, related to the surface energy balance at the atmosphere-land or atmosphere-ocean interface, to be evaluated.

2. Data and Methods

2.1. Model

2.1.1. Model and Experiment Description

ECMWF produces global weather forecasts from medium-range through to subseasonal and seasonal time scales. The deterministic 10-day high-resolution forecasts (HRES hereafter) are performed at 9 km horizontal resolution with 137 vertical levels (with 9 in the lowest 250 m). The ensemble 15-day forecasts (ENS) are performed at 18 km horizontal resolution with 91 vertical levels (with 6 in the lowest 250 m). However, testing of new model developments, as in this study, is often done at the lower resolution of 30 km and 137 vertical levels. In this study we use the experiments performed at this resolution by Arduini et al. (2019) with the single (SL) and multilayer (ML) snow schemes. These experiments were performed using Integrated Forecast System (IFS) Cycle 45r1, which was used operationally at ECMWF from July 2018 to June 2019.

The model uses a cubic octahedral Gaussian grid in the horizontal domain and the resolutions stated above are the approximate equivalent resolution in gridpoint space.

A set of 10-day coupled forecasts, initialized at 00UTC each day for the periods December–February 2013/2014 and 2017/2018 were performed with each version of the model. The atmospheric fields are initialized using the ECMWF operational analysis. Testing of new model developments at ECMWF is usually done by running forecasts for a recent period, in this case 2017/2018. However, a full complement of surface energy budget terms at Summit, Greenland was only available from July 2013 to June 2014, so an additional experiment was run for the winter of 2013/2014 to allow evaluation against these data (Miller et al., 2017). Since the ECMWF land surface analysis does not yet include snow parameters on multiple layers, the surface fields of the SL and ML coupled forecasts are initialized from global uncoupled (offline) simulations using the SL and ML snow schemes, respectively. These offline simulations cover the time period from June 2010 to June 2018 and were forced using reanalysis atmospheric data. The uncoupled single layer simulation produces initial conditions for the coupled forecasts with a single snow temperature, density, and total snow mass value for each gridbox. The uncoupled multilayer model produces liquid water content as an additional prognostic variable and produces all variables on each of the five layers. Further details of the initialization and experimental design may be found in Arduini et al. (2019).

In addition to the deterministic forecasts, two sets of 8-day ensemble forecasts with 21 members were also performed for the period December 2017 to February 2018 with the single-layer and the multilayer snow scheme, to demonstrate the impact of the new snow model on forecast reliability. The ensemble forecasts are initialized every day at 00UTC using the same procedure described for the deterministic forecasts. The horizontal resolution is about 30 km (TCo399) and 91 vertical levels are used. The ECMWF operational ensemble of data assimilation (EDA, Isaksen et al., 2010) and singular vector perturbations are used to take into account the uncertainty of the initial conditions. During each forecast integration, a Stochastically Perturbed Parameterization Tendencies (SPPT) scheme is used to take into account the uncertainties in the model formulation (Leutbecher et al., 2017). The number of simulated days in the ensemble forecasts is different from the deterministic ones to reduce the computational cost of these simulations.

In the model, turbulent fluxes are calculated within the surface layer, acting between the lowest atmospheric model level (~10 m) and the surface according to

$$\tau = \rho C_M U_{10m}^2 \tag{2}$$

$$SHF = \rho C_H U_{10m} \left(\theta_{10m} - \theta_{sfc} \right) \tag{3}$$

The transfer coefficients (C_M and C_H), needed to compute the surface stress (τ) and the sensible heat flux (*SHF*), are based on Monin Obukhov (M-O) similarity theory. They are a function of the roughness lengths of momentum/heat (z_{oM}/z_{oH}), and the bulk Richardson number (Ri_b). In the algorithm, the bulk Richardson number is first converted to the Obukhov length and then the Beljaars and Holtslag (1991) functions are used to compute the transfer coefficients.

The atmospheric model is coupled to the land surface model (HTESSEL, Balsamo et al., 2009), using the implicit scheme proposed by Best et al. (2004). In this coupling, the atmosphere and land are separated at the lowest model level and the atmospheric surface layer is considered to be part of the land surface scheme (Beljaars et al., 2018). Surface heterogeneity is reflected by a tile structure in HTESSEL and the energy balance is solved on each tile separately, using appropriate parameters for each surface type, but for each gridbox only a single aggregated value for each flux (weighted by the fraction of the gridbox area taken up by each tile) is seen by the atmosphere. The heat flux into the surface (ground heat flux, GHF) is calculated for each tile according to

$$GHF = \Lambda \left(T_{sfc} - T_{sn} \right) \tag{4}$$

where for the exposed snow tile, T_{sfc} is the temperature of the snow surface and T_{sn} is the temperature of the snowpack (top snow layer temperature in the ML scheme) and Λ is a surface conductivity parameter, which can be thought of as the thermal conductivity between the middle point of the top snow layer and the surface in the case of snow accumulating over bare soil or grass. HTESSEL uses two tiles for snow, one for exposed snow on low vegetation and one for snow under high vegetation. For the former, Λ is set to 7 W m^{-2} in both experiments, whereas for the high vegetation tile, Λ varies as a function of the snow water equivalent and density, following Beljaars et al. (2017), to ensure numerical stability in the case of very thin snow layers. The range of values of Λ for the high vegetation tile is between 9 W m^{-2} (for thin, low density snow) and 15 W m^{-2}. The fraction of each grid box covered by each tile type is derived from the global land cover characterization data set (GLCC vn 1.2, Loveland et al., 2000), combined with snow mass. The agregated value of *GHF*, across the two snow tiles, is passed to the snow model, to evolve the snow thermodynamics and mass.

The 2 m temperature is calculated diagnostically, as a weighted function of the temperature of the lowest model level, and the surface temperature of the low-vegetation tile as is consistent with surroundings at stations of the synoptic observing network. The model gridbox for Summit is 100% snow, but at Sodankylä the gridbox is a mixture (snow on low vegetation: 10%, snow under high vegetation: 89% and lake: 1% during this period).

The current snow scheme used in operational forecasts at ECMWF and included in HTESSEL uses energy conservation to describe the temporal evolution of the heat content and mass conservation driven by snowfall and melt to evolve snow mass. The description and evaluation of the current single layer snow model used in the IFS is reported by Dutra et al. (2010). The main processes and parameterizations are as follows: snow density is a prognostic field and varies due to overburden and thermal metamorphism (Anderson, 1976), as well as due to melt water retained in the snowpack (Lynch-Stieglitz, 1994). The liquid water content is diagnosed based on snow temperature at each time step. This enables also the rainfall interception by the snowpack to be taken into account. Snow albedo follows the empirical parameterization by Douville et al. (1995). The gridbox snow cover fraction is parameterized as a function of snow depth, varying linearly with snow depth between snow-free and fully snow-covered.

2.1.2. Changes to the Snow Scheme

The main difference in the new snow scheme compared to the current scheme is that it represents the vertical structure and temporal evolution of prognostic snow variables (i.e., temperature, density, and liquid water content) with multiple layers, rather than using a single layer for the whole snowpack. The new model uses the same parameterizations of snow albedo (both for exposed and forest snow) and snow cover fraction as the current operational model. An earlier version of this scheme, implemented in the EC-EARTH climate model, is described by Dutra et al. (2012) and tested in long climate simulations. In the multilayer formulation, the number of active snow layers and their thicknesses are computed diagnostically at the beginning of each time step before the prognostic snow fields are updated. The number of active layers (N) varies depending on the snow depth D_{sn}. For thin snow, a minimum number of one active layer is used, and for thick snow a maximum (N_{max}) of five layers are used. For a thick snowpack, the layer $N_{max} - 1$ (the penultimate layer from the bottom) is used as an accumulation layer, enabling a relatively high vertical resolution to be maintained at the interfaces with the atmosphere above and the soil underneath. An idealized example of the vertical discretization of a 1.0-m-thick snowpack is shown in Arduini et al. (2019), Figure 1). Liquid water content is also computed prognosticaly in the multilayer model, compared to the previous scheme where it was computed diagnostically based on snow temperature.

In addition to the multilayer formulation several additional parameterizations are included in the new model. (I) The heat conductivity is parameterized using the formulation of Calonne et al. (2011), taking into account water vapor diffusion effects, following Sun et al. (1999); (II) Transmission of solar radiation into the snow decreases exponentially with depth and is parameterized using a formulation adapted from Jordan (1991); (III) Density variations due to wind transport (snowdrift) are taken into account, in addition to the other compaction processes. This can be particularly effective for polar snow, for which snow temperature is extremely low throughout the winter and compaction due to other processes is limited (Brun et al., 1997; Decharme et al., 2016). Wind-driven compaction is parameterized using a mobility index combined with a wind-driven compaction index, following Decharme et al. (2016). (IV) The basal heat resistance is computed using a new physical formulation using the snow and soil thermal conductivities. Further details of the scheme can be found in Arduini et al. (2019).

2.2. Observational Data

In this study we make use of data from Sodankylä, Finland, and Summit, Greenland, which reside in different climate zones. Sodankylä is classified as continental sub-Arctic or boreal taiga, according to the Köppen land-type classification, whereas Summit station is located on an ice sheet. However, both Sodankylä, which has a seasonal snow pack with a maximum depth of around 80 cm, and Summit, which resides in the ice sheet's accumulation zone, are sites where forecasts are expected to benefit from an increased vertical resolution in the snowpack model. A common set of atmosphere and snow parameters are also measured at each site, enabling the same diagnostic analysis to be performed at both. This makes these suitable sites to conduct process-based evaluation of the new snow component for the IFS.

Upwelling and downwelling components of longwave (LW) and shortwave (SW) radiation are measured directly at both sites using pyrgeometers. At both sites the surface temperature was calculated according to

$$T_{sfc} = [(LW\uparrow - (1-\epsilon)LW\downarrow)/(\epsilon\sigma)]^{0.25} \tag{5}$$

where $\epsilon(= 0.985)$ is the surface emissivity (of fresh snow: Oke, 1987; Persson et al., 2002) and σ is the Stefan-Boltzmann constant.

At Sodankylä, the sensible and latent heat fluxes are measured at the micrometeorological mast by the eddy covariance method, using a three-axis sonic anemometer/thermometer, which provides direct measurements of the fluxes (Kangas et al., 2016). At Summit, due to a limited availability of fluxes from the eddy covariance method (Miller et al., 2017), the SHF and LHF are primarily calculated from temperature, wind, and humidity via the bulk aerodynamic method (Persson et al., 2002) and the two-level profile method (Steffen & Demaria, 1996). An important distinction between the sites is that Summit is very homogeneous, so M-O similarity theory is a suitable framework; however, the Sodankylä site is a mixture of open and forested terrain, where the applicability of similarity theory is questionable.

At Sodankylä, the ground heat flux (GHF), or atmosphere-snow heat flux is calculated as the sum of the conductive heat flux at a depth of 20 cm (*CHF*) and the heat flux convergence (*HFC*) in the top 20 cm of snow. This CHF is calculated according to

$$CHF = -k_{eff}\frac{\partial T}{\partial z} \tag{6}$$

where the temperature gradient is calculated from subsurface snow temperature observations. At Sodankylä, weekly snow density profiles (Leppänen et al., 2016), were interpolated in time and converted into an effective snow conductivity, k_{eff}, according to Sturm et al. (1997). The HFC is calculated according to

$$HFC = -c_{ice}\rho \times \frac{1}{2}\left[\frac{\partial T_{sfc}}{\partial t} + \frac{\partial T_{20cm}}{\partial t}\right](0.2) \tag{7}$$

where c_{ice} is the specific heat capacity of ice, ρ is the average density of the top 20 cm of the snow, and the temperature increments are calculated from hourly resolution observations. The equivalent fluxes at Summit were calculated by Miller et al. (2017). The procedure used to calculate these fluxes at Summit is subtly different, accounting for the fact that snow-temperature array is sinking over time due to the almost monotonic accumulation of snow-mass, whereas the snow-temperature array at Sodankylä is fixed with respect to the soil-snow interface.

The winter 2013–2014 period was chosen due to the availability of measurements of all SEB components at Summit, as well as Sodankylä. Further details of the Summit dataset, for this period, can be found in Miller et al. (2017). A detailed overview of the Sodankylä observatory, site specifics and collection methods may be found in Leppänen et al. (2016) for details of the manual snow observations, Essery et al. (2016) for details of automatic snow meteorological observations, and Kangas et al. (2016) for details of the atmospheric vertical profiles and turbulent fluxes. Note that at Sodankylä the radiation measurements (taken from the Radiometer Tower), are not precisely collocated with the turbulence (taken at the met tower) or the snow temperatures and density used to calculate the GHF (taken at the Intensive Observing Area).

2.3. Process-Oriented Diagnostics

The diagnostics used here to evaluate model improvements are based on those presented by Miller et al. (2018). Their motivation to separate the surface energy budget into a "driving term" ($LW\downarrow + SW_{net}$)

and “response terms” (*SHF, LHF, GHF,* and *-LW*↑) can be easily seen in observations from Arctic winter, where it is well known that boundary-layer and surface energy budget regimes are primarily driven by variations in *LW*↓, associated with synoptic scale variability in air mass properties (Miller et al., 2017; Pithan et al., 2014; Stramler et al., 2011). This type of behavior is illustrated in Figure 1, which shows the transition from cloudy conditions to cloud-free conditions at Sodankylä, Finland, in January 2014. During this period, clouds containing liquid water give way to clear sky conditions. The subsequent reduction in *LW*↓ results in a dramatic cooling at the surface (*a* ~ 30°C drop in surface temperature, T_{sfc}, and ~20°C drop in T_{2m} in 2 days) and a strong surface-based temperature inversion ($T_{sfc} < T_{2m}$). The radiative imbalance between the downwelling and upwelling longwave radiation in the cloud-free regime is compensated by the *SHF* and *GHF* terms, which both increase in response to the cooling of the surface.

The relationship between the driving term and each response term can be summarized with regression coefficients, for example, for the *SHF*:

$$SHF = \alpha_{SHF}(LW{\downarrow}+SW_{net}) + \beta_{SHF} \tag{8}$$

where each of the α's can be interpreted as a coupling strength parameter between the driving term and each response term. $LW{\downarrow} + SW_{net}$ is used instead of the total net radiation because it has no explicit dependence on the surface temperature (through $LW{\uparrow}$) from the driving term.

By substituting the right-hand side of these equations into Equation 1 one can derive the following expression relating the α's:

$$-1 = \alpha_{SHF} + \alpha_{LHF} + \alpha_{GHF} + \alpha_{-LW\uparrow} + \epsilon \tag{9}$$

where ϵ is the sum of the β terms divided by the driving term. Presenting them like this makes is clear that the α's provide direct information on the proportional response of each flux term, expressed as a fraction of the total change in radiative forcing. From this one can see that if, for example, the coupling to the land surface and the atmosphere is too strong in the model (i.e., $|\alpha_{GHF_{mod}} + \alpha_{SHF_{mod}} + \alpha_{LHF_{mod}}| < |\alpha_{GHF_{obs}} + \alpha_{SHF_{obs}} + \alpha_{LHF_{obs}}|$) then $|\alpha_{-LW\uparrow}|$, that is, surface temperature response, will be too weak and vice versa. Similarly, compensating errors in the strength of the coupling to the atmosphere ($\alpha_{SHF_{mod}} + \alpha_{LHF_{mod}}$) and coupling to the land surface ($\alpha_{GHF_{mod}}$) could result in the right surface-temperature response (i.e., correct $\alpha_{LW\uparrow}$), but for the wrong reasons.

Splitting the SEB into driving and response terms, and looking at process relationships in this way, has the desirable property that deficiencies in the behavior of the SEB can be diagnosed in isolation without the confounding effects of other sources of error, such as systematic or random cloud radiative forcing error, which are included in the ‘driving term’. In other words, one can assess whether the response to the radiative forcing is correct, irrespective of whether the forcing is itself correct.

In this framework, one could define the perfect model, as one who's α's are statistically indistinguishable from those derived from observations. One way to objectively determine if a linear regression coefficient in the model, α_{mod}, is significantly different to that of the observations, α_{obs}, is to use the test statistic, z, computed as the difference between the two regression coefficients divided by the standard error of the difference between the regression coefficients:

$$z = \frac{\alpha_{mod} - \alpha_{obs}}{S_{\alpha_{mod} - \alpha_{obs}}}, \tag{10}$$

where $S_{\alpha_{mod} - \alpha_{obs}} = \sqrt{S^2_{\alpha_{mod}} + S^2_{\alpha_{obs}}}$, $S^2_{\alpha} = \frac{1}{n-2}\frac{\sum(y-y')^2}{\sum(x-\bar{x})^2}$, y is the model or observed “response” (such as *SHF*), y' is its value predicted by the regression, x is the modeled or observed “driver” (such as $LW{\downarrow} + SW_{net}$), and $\bar{x}$ is its mean value. Under the null hypothesis ($\alpha_{mod} - \alpha_{obs} = 0$) z has a normal distribution and so can be used to test this hypothesis (Andrade & Estévez-Pérez, 2014).

The absolute value of z, defined above, provides a useful process-oriented metric of model performance, with smaller values of z indicating a better fit to observations. This complements the existing skill scores for near-surface weather parameters, generally used for evaluating changes to the forecasting system, which are

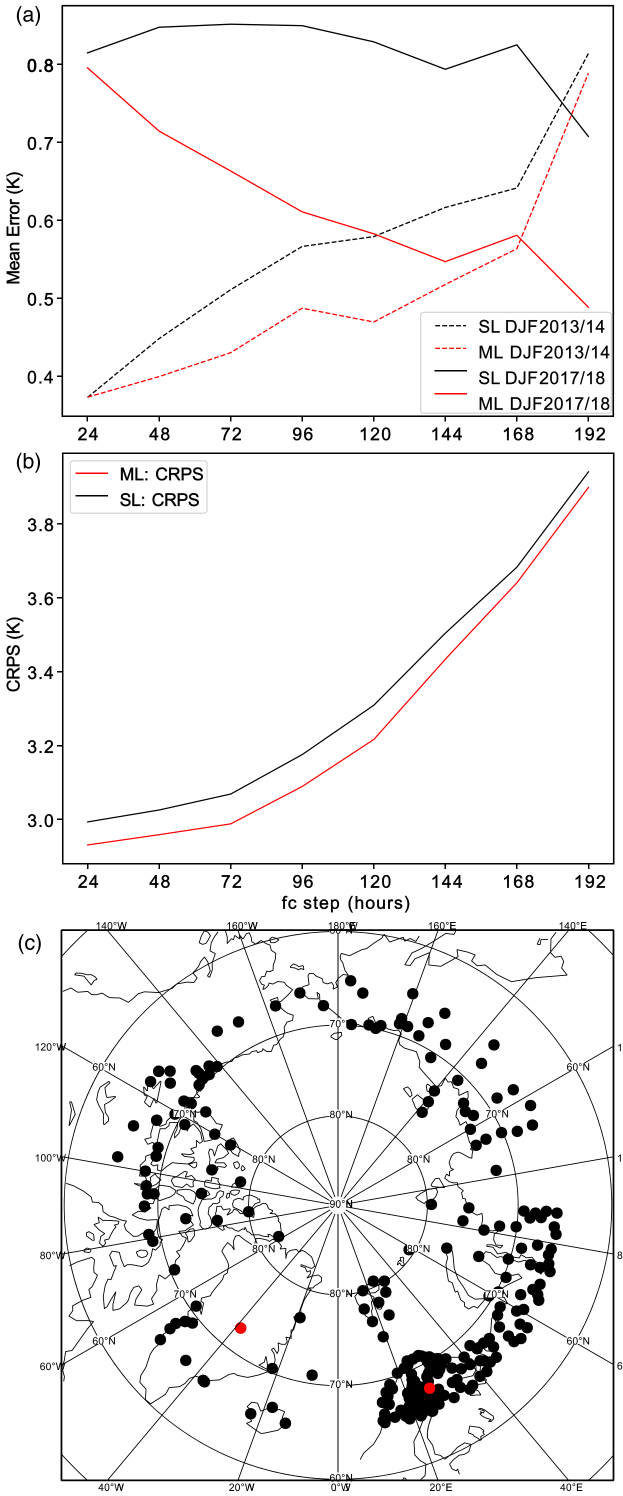

Figure 2. The 00 UTC 2 m temperature mean error for deterministic forecasts for DJF 2013/2014 and DJF 2017/2018 (a) and 2 m temperature continuous ranked probability score for 8-day ensemble forecasts for DJF 2017/2018 (CRPS; b) for the Arctic region (>65°N), compared to SYNOP stations (shown in c). Forecasts with single layer snow are shown in blue and multilayer snow are shown in red. The location of Sodankyla and summit stations are highlighted in the map by red dots.

typically based on the conventional weather stations and therefore limited to a few parameters such as total precipitation, 2 m temperature and humidity, 10 m wind and cloud cover.

3. Results

3.1. Evaluation Against Conventional Weather Stations

An anticipated outcome using the multilayer instead of the single-layer snow scheme is a reduction in the mean error of 2 m temperature forecasts over snow-covered surfaces. An evaluation of the change in 2 m temperature forecast skill between the two model formulations against SYNOP stations is performed over the Arctic region (above 65°N). There is a clear reduction in the winter warm bias when moving from the single layer control to multi-layer snow (Figure 2a) as well as a clear reduction in the *Continuous Ranked Probability Score in ENS forecasts (CRPS;* Figure 2b) at all lead times. Spatial maps of the change in mean-bias at Day 2 show a uniform reduction in temperature around the Arctic region, improving the mean error (see Figure 12 of Arduini et al., 2019). The fraction of grid-cells in midlatitudes with values of the CRPS > 5 K for 2 m temperature at a lead time of 5 days is one of ECMWF's headline scores, which are the set of scores used at ECMWF to evaluate long-term trends in forecast performance. Using the ML snow scheme results in a ~10% reduction in this metric in the Arctic (not shown), which is a large improvement in skill compared to other recent operational upgrades.

3.2. Evaluation at Supersites

3.2.1. Site Representativeness

For process-based evaluation at supersites to be informative in terms of the model performance at a regional level, it is important that the chosen sites are representative of the wider region of interest. Consistent with the Arctic-wide warm bias (Figures 2a, 3a, and 4a), 2 m temperature forecasts with the SL model exhibit a warm bias of 1.7°C at both Sodankylä and Summit, with the bias being largest for coldest temperatures. Atlaskin and Vihma (2012) present a multicenter analysis for northern Europe that shows that this warm bias at cold temperatures is characteristic of the wider region, common across a number of NWP models, and has been a long-standing error in ECMWF forecasts. Although, Sodankylä is a very heterogenous site, predominantly forested with pine trees (about 15 m tall) interspersed with clearings, verification against 2 m temperature observed at various locations across the station, including open and forested sites, show very similar error characteristics (Figure S2).

The inclusion of the multilayer snow reduces the 2 m temperature warm bias that is present during the coldest conditions at both sites (Figures 3d, 4d cf. Figures 3a, 4a). The mean error for the lowest temperature quantile at Sodankyla reduces from 8.1°C to 7.1°C and from 7.1°C to 4.0°C at Summit. This is consistent with Figure 2 and with the spatial maps of Arduini et al. (2019), who found that the improvement was largest for minimum 2 m temperature values. This suggests that these sites are indeed representative of the wider Arctic region.

3.2.2. Partitioning Sources of 2 m Temperature Error

As $LW\downarrow + SW_{net}$ is a major driver of 2 m temperature, errors in 2 m temperature are either due to errors in the driving term itself, the relationship between $LW\downarrow + SW_{net}$ and 2 m temperature, or a combination of both (assuming that errors in advection are negligible). Mean errors in the

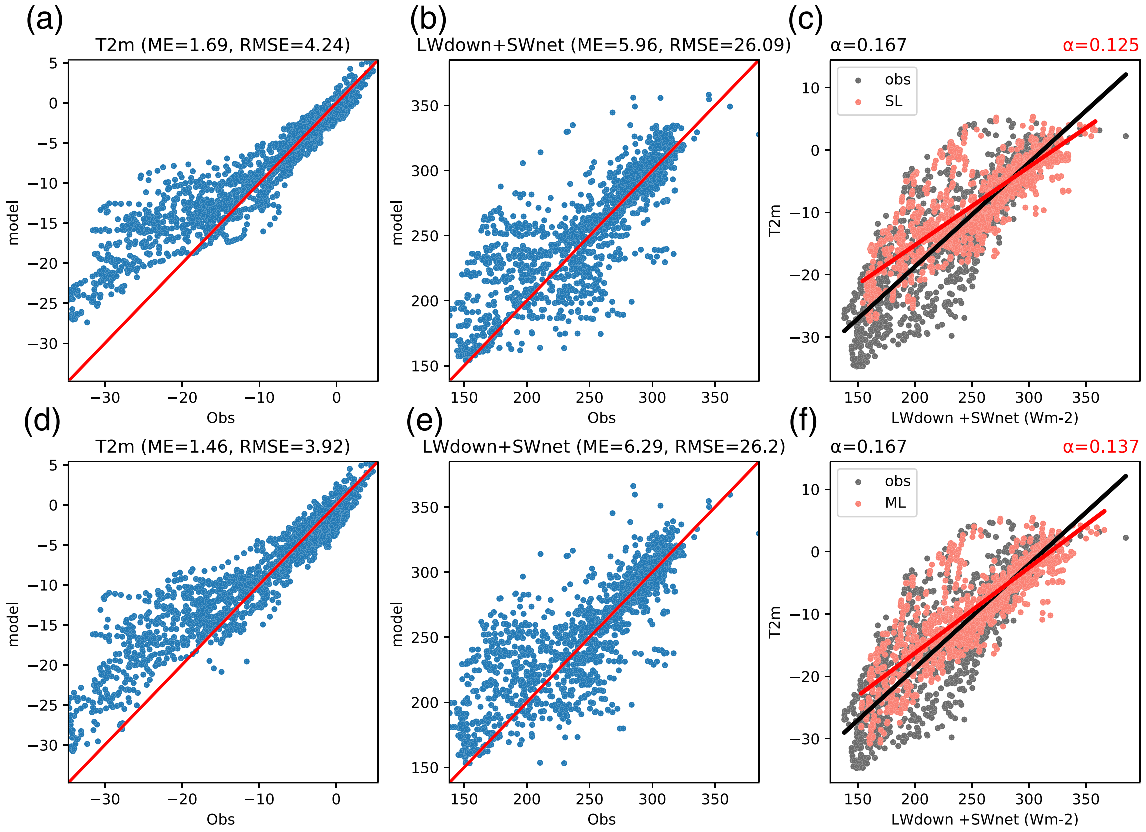

Figure 3. Hourly observed vs forecast (during Day 2) 2 m temperature (a and d), $LW\downarrow + SW_{net}$ (b and e), and the relationship between them (c and f) in observations (black) and each model formulation (red) for Sodankylä with single layer snow (top row) and multilayer snow (bottom row) for DJF 2013/2014. The regression coefficient is shown for the observations (black text) and the models (red text).

radiative forcing term are positive at Sodankylä (~6 W m^{-2}), particularly for low values of this term, and therefore contribute to the positive temperature errors (see Figure 3b). The mean error in the radiation term is negative at Summit (~8 W m^{-2}), shows that radiation errors are not responsible for the positive mean temperature bias there (see Figure 4b). In the absence of insolation, errors in the radiative forcing are likely to be associated with cloud radiative properties, such as the fraction of liquid water contained in Arctic clouds, which is a major driver of $LW\downarrow$ in the Arctic (Miller et al., 2017; Persson et al., 2017). Indeed, although the relationship between liquid-water path (LWP) and $LW\downarrow$ is quite well captured in the model, the forecasts, however, severely underestimate the LWP (Figure S3).

At both sites the 2 m temperature in the SL forecasts is less sensitive to changes in $LW\downarrow + SW_{net}$ than it is in observations (0.13°K/W m^{-2} compared to 0.17°K/W m^{-2} at Sodankylä and 0.14°K/W m^{-2} compared to 0.19°K/W m^{-2} at Summit). As a result, given the correct radiative forcing, the rate of change in temperature in a forecast at both sites will only be around three quarters of what it should be. The inclusion of the multilayer snow increases the sensitivity of 2 m temperature to radiative forcing at both sites. The lack of any substantial change in the driving term at either site (Figures 3e and 4e cf. Figures 3b and 4b) suggests that the reduction in T_{2m} error is due to this improvement in the response of 2 m temperature to radiative forcing. At low values of the $LW\downarrow + SW_{net}$ the values of 2 m temperature are lower for the ML experiment, which goes hand in hand with improved forecasts of cold conditions. The sensitivity at Summit is much improved, although slightly too high (0.20°K/W m^{-2}, see Table 2) in the ML experiment and improved but slightly too low at Sodankylä (0.14°K/W m^{-2}, see Table 1).

3.2.3. Surface Energy Budget Process Relationships

The sensitivity of 2 m temperature to radiative forcing is closely related to the sensitivity of the surface temperature. Indeed, the surface-temperature-$LW\downarrow + SW_{net}$ diagrams closely resemble those for 2 m

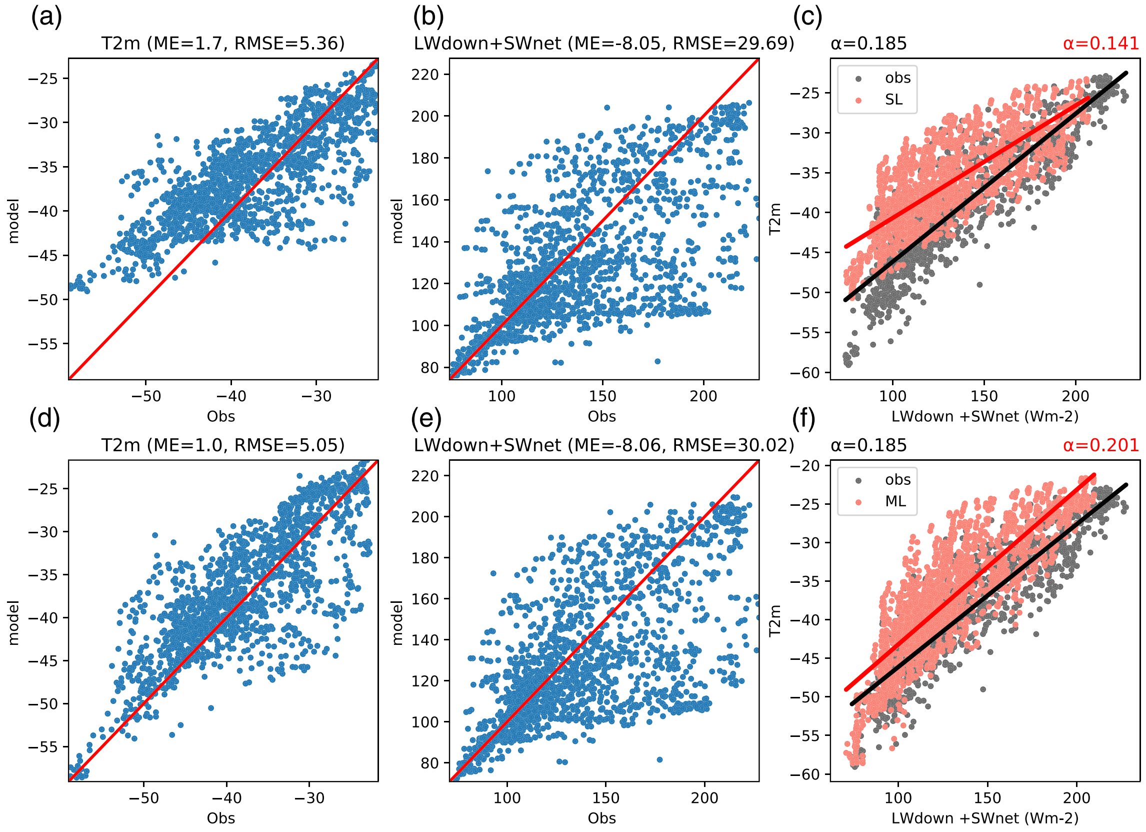

Figure 4. Hourly observed versus forecast (during Day 2) 2 m temperature (a and d), $LW\downarrow + SW_{net}$ (b and e), and the relationship between them (c and f) in observations (black) and each model formulation (red) for summit with single layer snow (top row) and multilayer snow (bottom row) for DJF 2013/20-14. The regression coefficient is shown for the observations (black text) and the models (red text).

temperature (Figures 3a and 4a cf. Figures 5a and 6a). Surface temperature is too insensitive to variations in the radiative forcing in the SL forecasts at both sites: 0.20°K/W m^{-2} compared to 0.13°K/W m^{-2} at Sodankyla and 0.24°K/W m^{-2} compared to 0.17°K/W m^{-2} at Summit (Figures 5a and 6a). This sensitivity increases at both sites in the ML forecasts but remains too low at Sodankylä (0.14°K/W m^{-2} Figure 5d) and becomes too high at Summit (0.27°K/W m^{-2} Figure 6d).

Because the energy budget is closed, an under or overly sensitive surface temperature (or $LW\uparrow$ equivalently) response to radiative forcing must be due to an error in the sensitivity of the remaining response terms (*SHF*, *LHF*, or *GHF*), as measured by α_{SHF}, α_{LHF} or α_{GHF}. By comparing these responses in the model to the response in observations we can understand the causes of systematic errors in the surface temperature sensitivity, and how this changes between model versions, from a process perspective.

Table 1
Observed and Modeled Regression Parameters at Sodankylä

Regression parameter (z statistic)			
Parameter	Observations	SL	ML
T_{sfc}	0.20	0.131 ($z = -34.3$, $p = 0.00$)	0.140 ($z = \mathbf{-29.0}$, $p = 0.00$)
SHF	−0.048	−0.252 ($z = \mathbf{-14.6}$, $p = 0.00$)	−0.250 ($z = -14.7$, $p = 0.00$)
GHF	−0.074	−0.169 ($z = -10.5$, $p = 0.00$)	−0.105 ($z = \mathbf{-4.99}$, $p = 2.97\text{e-}7$)
LHF	−0.053	−0.028 ($z = 4.97$, $p = 3.41\text{e}-7$)	−0.033 ($z = \mathbf{4.39}$, $p = 5.51\text{e-}6$)
$-LW\uparrow$	−0.79	−0.55 ($z = -29.0$, $p = 0.00$)	−0.58 ($z = \mathbf{-24.7}$, $p = 0.00$)
T_{2m}	0.165	0.125 ($z = -16.3$, $p = 0.00$)	0.133 ($z = \mathbf{-12.8}$, $p = 0.00$)

Note. Bold values highlight which z score is better.

Table 2
Table of Observed and Model Regression Parameters at Summit Station

Regression parameter (z statistic)			
Parameter	Observations	SL	ML
T_{sfc}	0.235	0.173 ($z = -18.2$, $p = 0.0$)	0.268 ($z = \mathbf{8.1}$, $p = 3.3e{-16}$)
SHF	−0.107	0.073 ($z = 14.5$, $p = 0.0$)	−0.128 ($z = \mathbf{-1.85}$, $p = 0.03$)
GHF	−0.366	−0.588 ($z = \mathbf{-8.3}$, $p = 0.0$)	−0.141 ($z = 8.99$, $p = 0.0$)
LHF	0.042	0.020 ($z = \mathbf{-15.5}$, $p = 0.0$)	0.003 ($z = -26.8$, $p = 0.0$)
$-LW\uparrow$	−0.666	−0.504 ($z = -17.1$, $p = 0.0$)	−0.760 ($z = \mathbf{8.5}$, $p = 0.0$)
T_{2m}	0.185	0.141 ($z = -13.2$, $p = 0.0$)	0.204 ($z = \mathbf{5.1}$, $p = 0.0$)

Note. Bold values highlight which z score is better.

To help in interpreting these PODs, it is useful to consider how the surface temperature response to radiative forcing depends on the turbulence regime (as defined by the Bulk-Richardson number, Ri) in observations (Figures S4 and S5). The surface-temperature sensitivity to radiative forcing is higher in nonturbulent regimes (0.21 K/W m^{-2} when $Ri > 0.25$) than in turbulent regimes (0.17 K/W m^{-2} when $Ri < 0.25$). This can be explained by the fact that in the turbulent regime, variations in radiative forcing can be balanced, to some extent, by variations in the turbulent heat fluxes (e.g., $\alpha_{SHF} = -0.13$ when $Ri > 0.25$). As Ri increases, the turbulent fluxes decrease and hence the fraction of incoming radiation they can balance decreases (e.g., $\alpha_{SHF} = -0.06$ when $Ri < 0.25$). The fraction balanced by LW↑ and GHF ($|\alpha_{GHF} + \alpha_{-LW\uparrow}|$) must therefore increase, allowing the surface temperature to become more responsive. This implies that a model with excessive turbulent diffusion in the atmosphere, for example, would have a surface-temperature sensitivity that was too low.

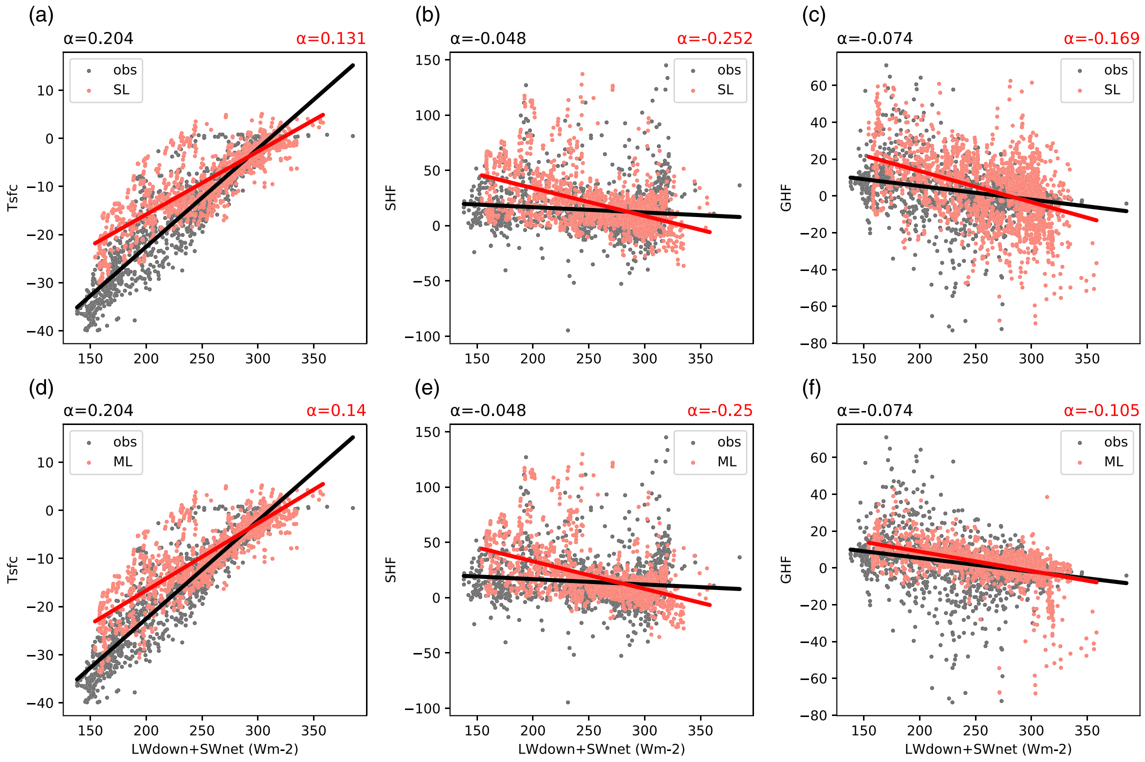

Figure 5. Process relationship diagrams and sensitivity parameters for surface temperature (T_{sfc}; left), sensible heat flux (SHF; middle), and ground heat flux (GHF; right) for Sodankyla, Finland. Observed values are shown in black, model values are shown in red for single layer snow (a–c) and multilayer snow (d–f). The line of best fit is shown for observations (gray line) and each model (pink line).

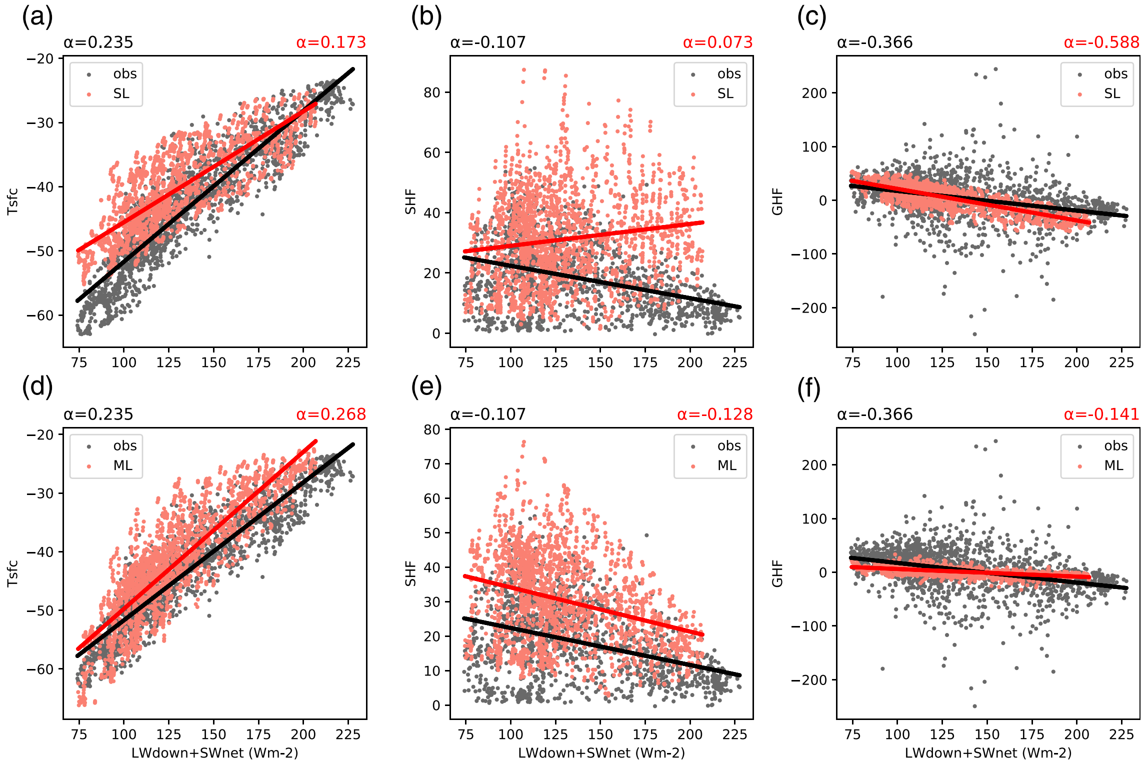

Figure 6. Process relationship diagrams and sensitivity parameters for surface temperature (T_{sfc}; left), sensible heat flux (*SHF*; middle), and ground heat flux (*GHF*; right) for summit, Greenland. Observed values are shown in black, model values are shown in red for single-layer snow (a–c) and multilayer snow (d–f). The line of best fit is shown for observations (gray line) and each model (pink line).

In the SL forecast the coupling strength to the land surface is too strong at both sites (i.e., $|\alpha_{GHF_{mod}}| > |\alpha_{GHF_{obs}}|$ see Figures 5c and 6c). The fraction of the radiative forcing going into heating the land surface is almost double what is observed at Sodankyla (−0.17 compared to −0.07) and 60% higher than observed at Summit (−0.59 compared to −0.37). The coupling to the atmosphere is also too high at Sodankyla (i.e., $|\alpha_{SHF_{mod}} + \alpha_{LHF_{mod}}| > |\alpha_{SHF_{obs}} + \alpha_{LHF_{obs}}|$ see Figures 5b and S6, and Tables 1 and 2), which also contributes to the surface temperature sensitivity being too low (i.e., $|\alpha_{-LW\uparrow_{mod}}| < |\alpha_{-LW\uparrow_{obs}}|$). At Summit the coupling to the atmosphere is too low (and $\alpha_{SHF_{mod}}$ even has the wrong sign, see Figure 6b) but because $|\alpha_{SHF_{mod}} + \alpha_{LHF_{mod}} + \alpha_{GHF_{mod}}|$ is too high overall (See Figures 6b, S7, and Table 2), the surface-temperature response is also too low, as it is at Sodankylä.

Using the multilayer instead of the single-layer snow scheme directly influences the coupling between the radiation and the *GHF*, that is, α_{GHF}, because the snow temperature (T_{sn}) used in the *GHF* calculation (Equation 4) is the temperature of a thin layer at the top of the snowpack rather than the snowpack's mean temperature. The temperature of the top layer is able to respond more rapidly to changes in radiative forcing than the snowpack mean temperature. As a result, there is effectively a decoupling of the deep snow layers from the atmosphere when moving from the SL to the ML scheme. This results in a reduction in the fraction of the radiative forcing which is balanced by the GHF (i.e., a reduction in $|\alpha_{GHF_{mod}}|$) at both sites (see Figures 5 and 6 and Tables 1 and 2). As a result, this leads to an increased and improved surface-temperature sensitivity at both sites. However, $|\alpha_{GHF_{mod}}|$ remains a bit too high at Sodankyla (−0.11 compared to −0.07 in observations), while it becomes too low at Summit (−0.14 compared to −0.37 in observations). The reduction in the magnitude of $\alpha_{GHF_{mod}}$ is also much larger at Summit (~20% of the original value) than at Sodankyla (~60% of the original value). This difference in the change is likely related to the deeper snowpack at Summit than at Sodankylä, but may also be related to the fact that the model gridbox at Sodankyla is mainly forest-covered and the coupling parameter, Λ (see Equation 4), for snow under forest is about 3 times that for exposed snow (~20 W m^{-2} compared to 7). As a result, a larger GHF will be maintained over the forested tile, compared to a

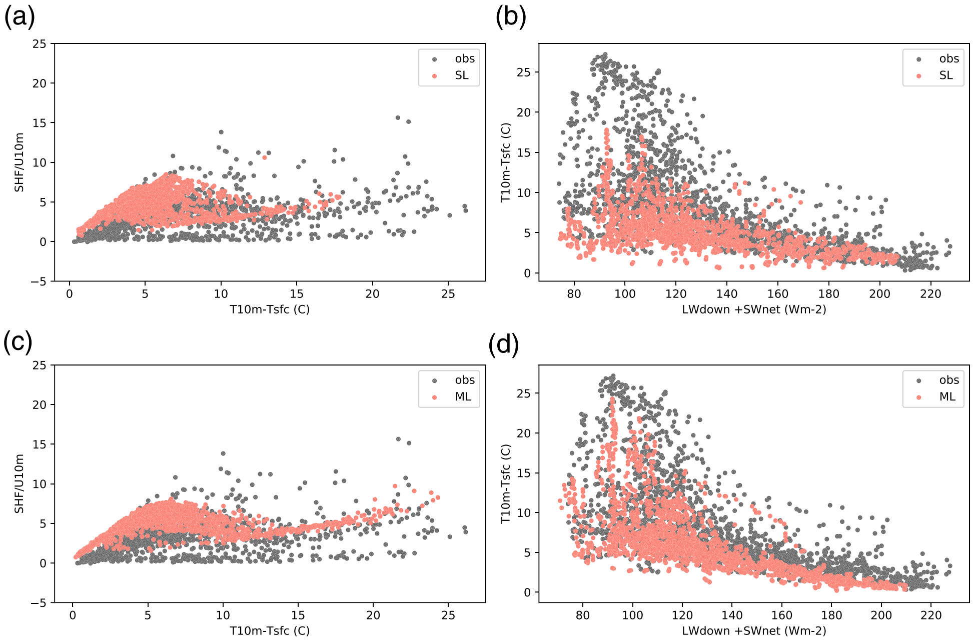

Figure 7. Sensible heat, scaled by wind speed, as a function of inversion strength at summit from forecasts with the single-lager model (SL, a) and multilayer model (ML, c). Inversion strength as a function of radiative forcing (LW↓ + $SW_{net:}$) for SL (b) and ML (d). Observations are shown in black and forecasts are shown in red.

case with lower Λ, therefore reducing the impact of the ML scheme on the gridbox mean surface temperature sensitivity.

Because the land and atmosphere represent a coupled system, the changes to the land surface parameterizations can also influence radiative and turbulent fluxes. For example, in the SL forecasts (and in ERA-Interim, see Miller et al., 2018) the sign and the magnitude of the response of SHF to the radiative forcing ($\alpha_{SHF_{mod}}$) at Summit is incorrect (0.07 compared to −0.11 in observations, Figure 6b and Table 2). Coupling to the multilayer snow changes the sign and magnitude, to a value of −0.13, bringing $\alpha_{SHF_{mod}}$ into close agreement with the observed value (Figure 6b and 6e). The response of the SHF improves because the ML version has more realistic inversion strength (T_{10m}-T_{sfc}) for a given value of incoming longwave (Figure 7b and 7d) which subsequently improves the distribution of SHF (Figure 7a and 7c) and its response to variations in radiative forcing.

The ability of a change in one of the model's parameterizations (in this case in the snow) to influence all surface energy fluxes is best highlighted and quantitatively measured by the differences of the SEB slope parameters. These should be used together to determine whether the simulation of the SEB has improved overall and to understand changes in the T_{sfc} sensitivity to variations in radiative forcing.

In contrast, improving the magnitude of $\alpha_{GHF_{mod}}$ at Sodankylä, does not result in a similar improvement in $\alpha_{SHF_{mod}}$ as at Summit. Instead, the SHF remains too responsive to variations in radiative forcing, and when a drop in incoming radiation cools the surface, the SHF increases too rapidly in response. As a result T_{sfc} still does not respond to variations in radiative forcing as much as in observations in the forecasts with ML snow. This suggests that another source of error exists outside the snow scheme.

4. Discussion

4.1. The Role of the Coupling to the Atmosphere

In the previous section, we showed that the coupling to the land surface was too strong in the SL simulations at both sites. The new snow model increased the response of the surface temperature by reducing the coupling to the land surface (i.e., $\alpha_{GHF_{mod}}$) in line with observations. However, at Sodankyla this was not sufficient to increase the surface-temperature sensitivity enough to match observations. This implies that the coupling to the atmosphere is too strong (also shown by the fact that $|\alpha_{SHF_{mod}} + \alpha_{LHF_{mod}}| > |\alpha_{SHF_{obs}} + \alpha_{LHF_{obs}}|$). This could either be because of errors in the formulation of the turbulent exchange in the surface layer (between 10 m and the surface) or in the outer layer (i.e., above 10 m). Errors associated with the large-scale dynamics or errors associated with boundary layer processes in adjacent areas could also provide an erroneous forcing on the boundary layer in the column above the site.

It is difficult to determine diagnostically which of these aspects is the culprit. In theory, one should be able to calculate the transfer coefficients in Equations 2 and 3, given both the observed flux and bulk properties at a given site (e.g., see Tjernström et al., 2005). In practice, however, in vegetated areas or complex terrain such as Sodankyla, the assumptions for M-O theory do not apply resulting in a large discrepancy between theory and practice. As a result, it is not always possible to evaluate the bulk transfer coefficients diagnostically. However, a positive wind speed bias at the lowest model level when low wind speeds are observed is a feature of both sites and will contribute to excessive turbulent fluxes at the surface during stable conditions (Figure S8).

Similarly, the turbulent exchange coefficients in the outer layer are hard to determine empirically and the current version of the IFS makes use of so-called "long-tail" stability functions for stable situations (Viterbo et al., 1999). These functions prescribe exchange coefficients which are larger, especially in strongly stable conditions ($Ri > 1$), than those prescribed by the M-O stability functions for stable situations (also known as "short-tail" functions). This choice was made to achieve an optimal performance in both the large-scale circulation and to avoid runway cooling near the surface (Sandu et al., 2013).

In an additional sensitivity study, the IFS was run with "short-tail" stability functions in stable boundary layers as well as with the new multilayer snow scheme. This reduces the fraction of radiation being balanced by the SHF,$|\alpha_{SHF}|$, and therefore increases, to some extent, the surface temperature sensitivity to radiative forcing at both sites compared to the ML-only runs (not shown). Such a change could not currently be implemented in the IFS globally without degrading synoptic forecast quality and increasing the near-surface cold bias over central and southern Europe (e.g., Sandu et al., 2013) but provides an example of a way in which the coupling strength to the atmosphere may be reduced, to bring $\alpha_{SHF_{mod}}$ into closer agreement with observed values at this site. Note that a reduction in the strength of $\alpha_{SHF_{mod}}$ could also be achieved by reducing the value of the bulk transfer coefficient for heat, C_H, in the surface layer (see Equation 3).

4.2. Other Applications of the Diagnostics

The current study has focused on understanding the impacts of a new snow model on the SEB at two Arctic sites during winter. This simplifies the analysis in two ways: first, as the Arctic is in perpetual night, errors in the surface albedo will not contribute to errors in the driving radiation term ($LW\downarrow + SW_{net}$). Second, energy going into melting snow, which is not directly measured, will be minimal during the period and as a result does not need to be considered in the analysis. If one were to extend the analysis to spring, this would not be the case. Errors in the prescription of the albedo would correspond to errors in the driving radiation term. In the presence of snowmelt, the additional term corresponding to the latent heat flux absorbed by the snowpack, would need to be included in the analysis. Despite complicating the analysis, as this term is difficult to measure, these details do not fundamentally change the interpretation of the diagnostics.

The methodology could also be extended to different climate zones, either to attribute sources of error or to look at the impact of changes in physical processes. For example, changes to the number of soil layers or other parameters, relevant to the land surface scheme or its coupling to the atmosphere, could be investigated at low-elevation sites in midlatitudes (such as Cabauw in the Netherlands or Lindenberg in Germany), which are usually snow free and have a long record of all the parameters used in the diagnostics. The only practical difference would be in the derivation of the GHF, particularly the conductive flux (at a depth of a few cm), which in this study is calculated from observed snow density and temperature. However, equivalent methods to calculate the GHF for snow-free soil, either using soil temperature in place of snow temperature, or using a heat flux plate, buried in the soil are well established (e.g., Liebethal & Foken, 2007).

4.3. Observations: Quality, Uncertainty, and Availability

It may also be relevant to consider what one can learn from these diagnostics, at sites where observations are limited. Armed with only observed radiative flux components and 2 m temperature one could determine if a systematic error in forecast 2 m or surface temperature was related to errors in the radiative forcing or to a systematic error in the temperature response, and further whether a given model change improved this relationship. However, without knowledge of the turbulent fluxes or the GHF, one would not be able to further interpret the reasons for this error in the temperature response, or evaluate whether a given change to the model physics was improving this response for the right reasons.

If in addition to the radiation components one has the terms required to calculate the GHF, but not data from a sonic anemometer to determine the SHF and LHF from the EC method, it is possible to estimate these fluxes from profiles of wind, temperature, and humidity using the bulk flux method (e.g., Persson et al., 2002). However, if the necessary bulk parameters are missing, inferring a missing term from the residual of the others is likely to lead to erroneous results due to the well-known energy balance closure problem, that is, that observed turbulent fluxes, using the EC method, tend to be lower than the available energy suggests they should be (see Foken, 2008 for a detailed explanation). Further, one should keep this closure issue in mind when comparing modeled and observed α_{SHF} and α_{LHF}.

An aspect not covered by this study, but of clear importance is the issue of measurement uncertainty. The uncertainty of a given radiation measurement is fairly small (~5 W m^{-2}), compared to the GHF components and turbulent heat fluxes (see Foken, 2008; Kohsiek et al., 2007; Miller et al., 2017). However, estimations of the random component of the error will not affect estimates of α, it is rather conditional bias that is of most concern. For example, if the underestimation of the observed *SHF*, mentioned above, is larger when $LW\downarrow + SW_{net}$ is high and smaller when $LW\downarrow + SW_{net}$ is low, then α_{SHF} could be underestimated. Understanding and accounting for this type of conditional error should be a priority for future work.

5. Conclusions

In this study we have presented a new way to evaluate model developments from the perspective of SEB process relationships for surface and 2 m temperature and the surface energy budget. These process-oriented diagnostics are applied to evaluate the impact of a new snow scheme in the ECMWF IFS at two Arctic sites: Summit station, in the center of the Greenland Ice Sheet and Sodankylä, a heterogeneous Arctic Taiga site in Finland. However, the use of these diagnostics is not restricted to snow-covered surfaces and they could be applied at any meteorological supersite to evaluate any relevant model change and ensure that any forecast improvements are occurring for the right reasons. The approach is shown to be complementary to, and useful for understanding the impact on, traditional skill scores computed against surface synoptic observations, which are more spatially abundant, but do not allow such detailed process analysis.

The approach we take is based on the idea that systematic errors in 2 m temperature can be partitioned into two distinct sources: errors in radiative forcing and errors in the response of surface and near-surface properties to variations in radiative forcing (i.e., $LW\downarrow + SW_{net}$, following Miller et al., 2018). It is shown that the weak response of 2 m and surface temperature to variations in radiative forcing is a common factor contributing to a warm bias (during cold conditions) in the operational forecasts produced at ECMWF for both sites and across the wider Arctic region.

Because the SEB is closed, systematic errors in the response of surface temperature to radiative forcing can be understood by analyzing the coupling strength between radiation and the energy balance terms, defined as the least squares regression parameter between the driving term: $LW\downarrow + SW_{net}$ and response terms: *SHF, LHF, GHF*, and $-LW\uparrow$. In the operational version of the IFS, which use a single-layer snow scheme, the total fraction of the radiative forcing balanced by the turbulent fluxes and ground heat flux is too high at both sites, as a result the fraction balanced by LW↑ (i.e., the surface temperature response) is too low. The coupling strength to the land surface is too strong due to the large thermal inertia associated with having to warm or cool the entire snowpack in the single-layer model.

Using a multilayer snow scheme results in an overall improvement in Arctic 2 m temperature forecasts, reducing a systematic warm bias, particularly during cold events. Improvements in the mean 2 m temperature biases at each site go hand in hand with an increased sensitivity of surface temperature to radiative forcing. Changing from the single-layer to the multi-layer scheme reduces the coupling strength between the radiation and the GHF directly, because the snow temperature used to calculate the GHF is the temperature of a thin layer at the top of the snowpack rather than the snowpack's mean temperature, which can respond faster (Equation 4). Subsequent changes in the coupling between the radiative forcing and the other SEB response terms (*SHF, LHF*, and *LW*↑) and ultimately T_{2m} occur indirectly, through the impact on surface-temperature, due to the tightly coupled nature of the land-atmosphere system. This is particularly noticeable in the results for Summit, Greenland where the response of the SHF, to changes in radiative forcing, markedly improves as an indirect response to improved land surface coupling. This is an interesting example of how interconnected the various model components are and hence the need to evaluate coupled behavior with such diagnostics.

The diagnostic framework provides a coupled perspective of the impact of a new model component, which goes beyond the evaluation of coupled forecasts in Arduini et al. (2019), and could be applied, in principle, to more detailed snow model process evaluation, which is often conducted in standalone model configurations forced by observations (e.g., Essery et al., 2009). Arctic winter provides a useful testing ground for the diagnostics shown here, since low levels of incoming shortwave radiation means that albedo can be ignored and SW penetration into the snow, which hinders estimation of heat transfer and heat content in the snow, is not an issue. Also, at this type of environment *LW*↓ is approximately balanced by *SHF, GHF* and *LW*↑ (SW and *LHF* terms are an order of magnitude smaller: Figure 1), simplifying the interpretation of the analysis. However, these diagnostics could be usefully applied to midlatitudes, for example, helping to diagnose sources of error in the diurnal cycle, where latent heat and coupling to the soil become more important (e.g., Panwar et al., 2019; Schmederer et al., 2019). An important next step would also be to link these diagnostics of the surface energy budget to diagnostics of boundary layer height (e.g., Lavers et al., 2019), whose growth is known to modulate the heating rates during the morning-leg of the diurnal cycle (e.g., Panwar et al., 2019).

Data Availability Statement

The Summit Greenland observed surface energy budget data set is available online in the National Science Foundation's Arctic Data Center. (Matthew Shupe and Nathaniel Miller. 2016. Surface energy budget at Summit, Greenland. NSF Arctic Data Center. https://arcticdata.io/catalog/view/doi:10.18739/A2Z37J). The Sodankyla surface energy budget data are available from Finnish Meteorological Institute (http://litdb.fmi.fi). Both are published under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. The forecasts with single-layer and multilayer snow model will be published at the Zenodo repository, following acceptance of the manuscript, with the following https://doi.org/10.5281/zenodo.3755373.

Acknowledgments

The work described in this article has received funding from the European Union's Horizon 2020 Research and Innovation program through Grant Agreement 727862 APPLICATE. The content of the article is the sole responsibility of the author(s) and it does not represent the opinion of the European Commission, and the Commission is not responsible for any use that might be made of information contained. The authors would like to thank Anna Kontu from the Finnish Meteorological Institute (FMI) for providing the snow pit measurements at Sodankylä. The staff at FMI-ARC are thanked for data collection and maintenance of instruments at the Sodankylä site. They would also like to thank Nate Miller (University of Wisconsin-Madison) and Matt Shupe (University of Colorado and NOAA Earth System Research Laboratories) for providing the data from Summit, Greenland.

References

Anderson, E. A. (1976). A point energy and mass balance model of a snow cover (Tech. Rep. No. NWS 19). National Oceanic and Atmospheric Administration (NOAA).

Andrade, J. M., & Estévez-Pérez, M. G. (2014). Statistical comparison of the slopes of two regression lines: A tutorial. *Analytica Chimica Acta*, *838*, 1–12. https://doi.org/10.1016/j.aca.2014.04.057

Arduini, G., Balsamo, G., Dutra, E., Day, J. J., Sandu, I., Boussetta, S., & Haiden, T. (2019). Impact of a multi-layer snow scheme on near-surface weather forecasts. *Journal of Advances in Modeling Earth Systems*, *11*, 4687–4710. https://doi.org/10.1029/2019MS001725

Atlaskin, E., & Vihma, T. (2012). Evaluation of NWP results for wintertime nocturnal boundary-layer temperatures over Europe and Finland. *Quarterly Journal of the Royal Meteorological Society*, *138*(667), 1440–1451. https://doi.org/10.1002/qj.1885

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M., & Betts, A. K. (2009). A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the integrated forecast system. *Journal of Hydrometeorology*, *10*(3), 623–643.

Bauer, P., Magnusson, L., Thépaut, J.-N., & Hamill, T. M. (2016). Aspects of ECMWF model performance in polar areas. *Quarterly Journal of the Royal Meteorological Society*, *142*(695), 583–596. https://doi.org/10.1002/qj.2449

Beljaars, A., Balsamo, G., Bechtold, P., Bozzo, A., Forbes, R., Hogan, R. J., et al. (2018). The numerics of physical parametrization in the ECMWF model. *Frontiers in Earth Science*, *6*. https://doi.org/10.3389/feart.2018.00137

Beljaars, A., Dutra, E., Balsamo, G., & Lemarié, F. (2017). On the numerical stability of surface–atmosphere coupling in weather and climate models. *Geoscientific Model Development*, *10*(2), 977–989. https://doi.org/10.5194/gmd-10-977-2017

Beljaars, A. C. M., & Holtslag, A. A. M. (1991). Flux parameterization over land surfaces for atmospheric models. *Journal of Applied Meteorology*, *30*(3), 327–341. https://doi.org/10.1175/1520-0450(1991)030<0327:FPOLSF>2.0.CO;2

Best, M. J., Beljaars, A., Polcher, J., & Viterbo, P. (2004). A proposed structure for coupling tiled surfaces with the planetary boundary layer. *Journal of Hydrometeorology*, *5*(6), 1271–1278. https://doi.org/10.1175/JHM-382.1

Brun, E., Martin, E., & Spiridonov, V. (1997). Coupling a multi-layered snow model with a GCM. *Annals of Glaciology*, *25*, 66–72. https://doi.org/10.3189/S0260305500013811

Calonne, N., Flin, F., Morin, S., Lesaffre, B., du Roscoat, S. R., & Geindreau, C. (2011). Numerical and experimental investigations of the effective thermal conductivity of snow. *Geophysical Research Letters*, *38*, L23501. https://doi.org/10.1029/2011GL049234

Cuxart, J., Holtslag, A. A. M., Beare, R. J., Bazile, E., Beljaars, A., Cheng, A., et al. (2006). Single-column model intercomparison for a stably stratified atmospheric boundary layer. *Boundary-Layer Meteorology*, *118*(2), 273–303. https://doi.org/10.1007/s10546-005-3780-1

Decharme, B., Brun, E., Boone, A., Delire, C., Le Moigne, P., & Morin, S. (2016). Impacts of snow and organic soils parameterization on northern Eurasian soil temperature profiles simulated by the ISBA land surface model. *The Cryosphere*, *10*(2), 853–877. https://doi.org/10.5194/tc-10-853-2016

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Douville, H., Royer, J.-F., & Mahfouf, J.-F. (1995). A new snow parameterization for the Meteo-France climate model. *Climate Dynamics*, *12*(1), 21–35.

Dutra, E., Balsamo, G., Viterbo, P., Miranda, P. M., Beljaars, A., Schär, C., & Elder, K. (2010). An improved snow scheme for the ECMWF land surface model: Description and offline validation. *Journal of Hydrometeorology*, *11*(4).

Dutra, E., Viterbo, P., Miranda, P. M., & Balsamo, G. (2012). Complexity of snow schemes in a climate model and its impact on surface energy and hydrology. *Journal of Hydrometeorology*, *13*(2), 521–538. https://doi.org/10.1175/JHM-D-11-072.1

Essery, R. (2010). Snow parameterisation in GCMs. In R. L. Armstrong & E. Brun (Eds.), *Snow and Climate* (pp. 145–156). Cambridge: Cambridge University Press.

Essery, R., Kontu, A., Lemmetyinen, J., Dumont, M., & Ménard, C. B. (2016). A 7-year dataset for driving and evaluating snow models at an Arctic site (Sodankylä, Finland). *Geoscientific Instrumentation, Methods and Data Systems*, *5*(1), 219–227. https://doi.org/10.5194/gi-5-219-2016

Essery, R., Rutter, N., Pomeroy, J., Baxter, R., Stähli, M., Gustafsson, D., et al. (2009). SnowMIP2: An evalution of forest snow process simulation. *Bulletin of the American Meteorological Society*, *90*(8), 1120–1136. https://doi.org/10.1175/2009BAMS2629.1

Foken, T. (2008). The energy balance closure problem: An overview. *Ecological Applications*, *18*(6), 1351–1367. https://doi.org/10.1890/06-0922.1

Haiden, T., Sandu, I., Balsamo, G., Arduini, G., & Beljaars, A. (2018). Addressing biases in near-surface forecasts|ECMWF. *ECMWF Newsletter*, *157*, 20–25. https://doi.org/10.21957/eng71d53th

Holtslag, A. A. M., Svensson, G., Baas, P., Basu, S., Beare, B., Beljaars, A. C. M., et al. (2013). Stable atmospheric boundary layers and diurnal cycles: Challenges for weather and climate models. *Bulletin of the American Meteorological Society*, *94*(11), 1691–1706. https://doi.org/10.1175/BAMS-D-11-00187.1

Illingworth, A. J., Hogan, R. J., O'Connor, E., Bouniol, D., Brooks, M. E., Delanoé, J., et al. (2007). Cloudnet. *Bulletin of the American Meteorological Society*, *88*(6), 883–898. https://doi.org/10.1175/BAMS-88-6-883

Isaksen, L., Bonavita, M., Buizza, R., Fisher, M., Haseler, J., Leutbecher, M. & Raynaud, L. (2010) Ensemble of data assimilations at ECMWF. Research Department Technical Memorandum No. 636, ECMWF, Shinfield Park, Reading RG29AX, UK, (available online at: http://www.ecmwf.int/publications/).

Jordan, R. (1991). A one-dimensional temperature model for a snow cover: Technical documentation for SNTHERM (CRREL Special Rep. 91-b). Hanover, NH: Cold regions research and engineering lab.

Jung, T., Gordon, N. D., Bauer, P., Bromwich, D. H., Chevallier, M., Day, J. J., et al. (2016). Advancing polar prediction capabilities on daily to seasonal time scales. *Bulletin of the American Meteorological Society*, *97*(9), 1631–1647. https://doi.org/10.1175/BAMS-D-14-00246.1

Kangas, M., Rontu, L., Fortelius, C., Aurela, M., & Poikonen, A. (2016). Weather model verification using Sodankylä mast measurements. *Geoscientific Instrumentation, Methods and Data Systems*, *5*(1), 75–84. https://doi.org/10.5194/gi-5-75-2016

Kohsiek, W., Liebethal, C., Foken, T., Vogt, R., Oncley, S. P., Bernhofer, C., & Debruin, H. A. R. (2007). The Energy Balance Experiment EBEX-2000. Part III: Behaviour and quality of the radiation measurements. *Boundary-Layer Meteorology*, *123*(1), 55–75. https://doi.org/10.1007/s10546-006-9135-8

Koster, R. D., Dirmeyer, P. A., Guo, Z., Bonan, G., Chan, E., Cox, P., et al. (2004). Regions of strong coupling between soil moisture and precipitation. *Science*, *305*(5687), 1138–1140. https://doi.org/10.1126/science.1100217

Koster, R. D., Sud, Y. C., Guo, Z., Dirmeyer, P. A., Bonan, G., Oleson, K. W., et al. (2006). GLACE: The global land–atmosphere coupling experiment. Part I: Overview. *Journal of Hydrometeorology*, *7*(4), 590–610. https://doi.org/10.1175/JHM510.1

Lavers, D. A., Beljaars, A., Richardson, D. S., Rodwell, M. J., & Pappenberger, F. (2019). A forecast evaluation of planetary boundary layer height over the ocean. *Journal of Geophysical Research: Atmospheres*, *124*, 4975–4984. https://doi.org/10.1029/2019JD030454

Leppänen, L., Kontu, A., Sjöblom, H., & Pulliainen, J. (2016). Sodankylä manual snow survey program. *Geoscientific Instrumentation, Methods and Data Systems*, *5*(1), 163–179. https://doi.org/10.5194/gi-5-163-2016

Leutbecher, M., Lock, S. J., Ollinaho, P., Lang, S. T., Balsamo, G., Bechtold, P., et al. (2017). Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, *143*(707), 2315–2339. https://doi.org/10.1002/qj.3094

Liebethal, C., & Foken, T. (2007). Evaluation of six parameterization approaches for the ground heat flux. *Theoretical and Applied Climatology*, *88*(1-2), 43–56. https://doi.org/10.1007/s00704-005-0234-0

Lindvall, J., & Svensson, G. (2015). The diurnal temperature range in the CMIP5 models. *Climate Dynamics*, *44*(1), 405–421. https://doi.org/10.1007/s00382-014-2144-2

Loveland, T. R., Reed, B. C., Brown, J. F., Ohlen, D. O., Zhu, Z., Yang, L., & Merchant, J. W. (2000). Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. *International Journal of Remote Sensing*, *21*(6–7), 1303–1330. https://doi.org/10.1080/014311600210191

Lynch-Stieglitz, M. (1994). The development and validation of a simple snow model for the GISS GCM. *Journal of Climate*, *7*(12), 1842–1855. https://doi.org/10.1175/1520-0442(1994)007<1842:TDAVOA>2.0.CO;2

Ma, H.-Y., Klein, S. A., Xie, S., Zhang, C., Tang, S., Tang, Q., et al. (2018). CAUSES: On the role of surface energy budget errors to the warm surface air temperature error over the Central United States. *Journal of Geophysical Research: Atmospheres*, *123*, 2888–2909. https://doi.org/10.1002/2017JD027194

Miller, N. B., Shupe, M. D., Cox, C. J., Noone, D., Persson, P. O. G., & Steffen, K. (2017). Surface energy budget responses to radiative forcing at summit, Greenland. *The Cryosphere*, *11*(1), 497–516. https://doi.org/10.5194/tc-11-497-2017

Miller, N. B., Shupe, M. D., Lenaerts, J. T. M., Kay, J. E., de Boer, G., & Bennartz, R. (2018). Process-based model evaluation using surface energy budget observations in Central Greenland. *Journal of Geophysical Research: Atmospheres*, *123*, 4777–4796. https://doi.org/10.1029/2017JD027377

Oke, T. R. (1987). *Boundary layer climates* (2nd ed.). Abingdon, Oxfordshire: Routledge. https://doi.org/10.1017/CBO9781107415324.004

Panwar, A., Kleidon, A., & Renner, M. (2019). Do surface and air temperatures contain similar imprints of evaporative conditions? *Geophysical Research Letters*, *46*, 3802–3809. https://doi.org/10.1029/2019GL082248

Persson, P., Ola, G., Shupe, M. D., Perovich, D., & Solomon, A. (2017). Linking atmospheric synoptic transport, cloud phase, surface energy fluxes, and sea-ice growth: Observations of midwinter SHEBA conditions. *Climate Dynamics*, *49*(4), 1341–1364. https://doi.org/10.1007/s00382-016-3383-1

Persson, P., Olga, G., Fairall, C. W., Andreas, E. L., Guest, P. S., & Perovich, D. K. (2002). Measurements near the atmospheric surface flux group tower at SHEBA: Near-surface conditions and surface energy budget. *Journal of Geophysical Research*, *107*(C10), 8045. https://doi.org/10.1029/2000JC000705

Pithan, F., Medeiros, B., & Mauritsen, T. (2014). Mixed-phase clouds cause climate model biases in Arctic wintertime temperature inversions. *Climate Dynamics*, *43*(1–2), 289–303. https://doi.org/10.1007/s00382-013-1964-9

Sandu, I., Beljaars, A., Bechtold, P., Mauritsen, T., & Balsamo, G. (2013). Why is it so difficult to represent stably stratified conditions in numerical weather prediction (NWP) models? *Journal of Advances in Modeling Earth Systems*, *5*, 117–133. https://doi.org/10.1002/jame.20013

Santanello, J. A. Jr., Dirmeyer, P. A., Ferguson, C. R., Findell, K. L., Tawfik, A. B., Berg, A., et al. (2018). Land–atmosphere interactions: The LoCo perspective. *Bulletin of the American Meteorological Society*, *99*(6), 1253–1272. https://doi.org/10.1175/BAMS-D-17-0001.1

Schmederer, P., Sandu, I., Haiden, T., Beljaars, A., Leutbecher, M., Becker, C., et al. (2019). Use of super-site observations to evaluate near-surface temperature forecasts. ECMWF newsletter number 161—Autumn 2019. https://doi.org/10.21957/fa518ps439

Steffen, K., & Demaria, T. (1996). Surface energy fluxes of Arctic winter sea ice in Barrow Strait. *Journal of Applied Meteorology*, *35*(11), 2067–2079. https://doi.org/10.1175/1520-0450(1996)035<2067:SEFOAW>2.0.CO;2

Sterk, H. A. M., Steeneveld, G. J., & Holtslag, A. A. M. (2013). The role of snow-surface coupling, radiation, and turbulent mixing in modeling a stable boundary layer over Arctic Sea ice. *Journal of Geophysical Research: Atmospheres*, *118*, 1199–1217. https://doi.org/10.1002/jgrd.50158

Stramler, K., Del Genio, A. D., & Rossow, W. B. (2011). Synoptically driven Arctic winter states. *Journal of Climate*, *24*(6), 1747–1762. https://doi.org/10.1175/2010JCLI3817.1

Sturm, M., Holmgren, J., König, M., & Morris, K. (1997). The thermal conductivity of seasonal snow. *Journal of Glaciology*, *43*(143), 26–41. https://doi.org/10.3189/s0022143000002781

Sun, S., Jin, J., & Xue, Y. (1999). A simple snow-atmosphere-soil transfer model. *Journal of Geophysical Research*, *104*(D16), 19,587–19,597. https://doi.org/10.1029/1999JD900305

Svensson, G., Holtslag, A. A. M., Kumar, V., Mauritsen, T., Steeneveld, G. J., Angevine, W. M., et al. (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. *Boundary-Layer Meteorology*, *140*(2), 177–206. https://doi.org/10.1007/s10546-011-9611-7

Tjernström, M., Žagar, M., Svensson, G., Cassano, J. J., Pfeifer, S., Rinke, A., et al. (2005). Modelling the Arctic boundary layer: An evaluation of six Arcmip regional-scale models using data from the Sheba project. *Boundary-Layer Meteorology*, *117*(2), 337–381. https://doi.org/10.1007/s10546-004-7954-z

Viterbo, P., Beljaars, A., Mahfouf, J.-F., & Teixeira, J. (1999). The representation of soil moisture freezing and its impact on the stable boundary layer. *Quarterly Journal of the Royal Meteorological Society*, *125*(559), 2401–2426. https://doi.org/10.1002/qj.49712555904

Walters, D., Baran, A. J., Boutle, I., Brooks, M., Earnshaw, P., Edwards, J., et al. (2019). The Met Office unified model global atmosphere 7.0/7.1 and JULES global land 7.0 configurations. *Geoscientific Model Development*, *12*(5), 1909–1963. https://doi.org/10.5194/gmd-12-1909-2019

Working group on numerical experimentation (WGNE): Systematic error survey results summary (2019): https://www.wcrp-climate.org/JSC40/12.7b.%20WGNE_Systematic_Error_Survey_Results_20190211.pdf

Zadra, A., Williams, K., Frassoni, A., Rixen, M., Adames, Á. F., Berner, J., et al. (2018). Systematic errors in weather and climate models: Nature, origins, and ways forward. *Bulletin of the American Meteorological Society*, *99*(4), ES67–ES70. https://doi.org/10.1175/BAMS-D-17-0287.1