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

Corresponding author: Jonathan Day (jonathan.day@ecmwf.int)

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 multi-layer snow model improves this by reducing the coupling-strength to the land-surface.

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 2m temperature, are happening for the right reasons. Correctly capturing such process relationships is a necessary step towards 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). The multi-layer scheme is expected to replace a single layer snow scheme in the operational forecasting system, enhancing the 2m temperature forecast reliability and skill across the northern hemisphere in boreal winter.

Plain Language Summary

Predicting 2m-temperature on timescales from hours to decades ahead is of high importance to a wide range of end users. However, it is also extremely are difficult to get right due to the large number of processes involved. 2m-temperature is affected by a large number of atmospheric processes such as those related to turbulent mixing, radiation, cloud 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 that improvements between one forecast system release and the next have occurred for the correct reason. 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 ECMWF IFS.

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 2m temperature, at timescales from minutes to decades, as highlighted by a recent survey of modelling centres conducted by the World Meteorological Organisation's Working Group on Numerical Experimentation (WGNE, 2019). 2m 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).

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.

In this study we present a set of Process Oriented Diagnostics (PODs) designed to assess the response of the 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 the coupling with the underlying medium (i.e. soil or snow) or to errors in the strength of the 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 build on the ideas of Miller et al. (2018), are based on the idea that the surface energy budget:

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

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 towards 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.

The 'driving' of the SEB by radiative forcing 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 $LW\downarrow$, associated with synoptic scale variability in air-mass properties (Miller et al., 2017; Pithan et al., 2014; Stramler et al., 2011). This type of behaviour 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 $LW\downarrow$ 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 two 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. Such sharp transitions between the cloudy and clear-sky states leads to a bimodal frequency distribution in the $LW\downarrow$-inversion strength space (Fig S1, see also Pithan et al., 2014)

This case study also highlights the importance of thermodynamic coupling between the atmosphere and the snow. The cooling of the snowpack is largest and most rapid near the surface during the transition between regimes. 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 thermal insulation of the snowpack (Figure 1). The resultant gradient within the snow pack is important for determining the magnitude of the heat-flux within the snow.

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, such as seen in the case study above, cannot be captured. In particular, with a single layer of snow, 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 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). It is expected that the inclusion of a multi-layer snow-scheme will result in a more responsive surface 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.

Such a multi-layer snow scheme has recently been 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 2m 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. Supersites, such as Sodankylä, Finland, and Summit, Greenland on the other hand collect a much wider set of observations, so can be used to evaluate model changes from a process-oriented perspective.

In this study the PODs described above will be applied to the single-layer and multi-layer snow forecast experiments and compared with those derived from observations at these two sites in order to evaluate whether the improvements in 2m 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 behaviour of the surface-atmosphere interaction at those locations. The analysis builds on the techniques of Miller et al. (2018) and applies these to the supersite observations in order to evaluate models and guide the model development process.

Although the analysis focusses on the impact of a new snow model in the Arctic during winter, the suite of PODs presented in this paper could be applied to any site, with appropriate instrumentation, or any season, to evaluate the impact of any model change related to the atmosphere-land or atmosphere-ocean interface, in terms of processes related to the surface energy budget.

2 Data & Methods

2.1 Model

2.1.1 Model and Experiment Description

ECMWF produces global weather forecasts from medium-range through to sub-seasonal and seasonal timescales. The deterministic ten-day high-resolution forecasts (HRES hereafter) are performed at 9 km horizontal resolution with 137 vertical levels (with 9 in the lowest 250 metres). The ensemble 15-day forecasts (ENS) are performed at 18 km horizontal resolution with

91 vertical levels (with 6 in the lowest 250 metres). However, testing of new model developments, as in this study, is often done at the lower resolution of 30km 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 multi-layer (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, initialised at 00UTC each day for the period December-February 2013/14, were performed with each version of the model. The atmospheric fields are initialized using the ECMWF operational analysis. 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. Further details of the initialisation and experimental design may be found in Arduini et al. (2019).

In addition to the deterministic forecasts, two sets of 8-day ensemble coupled forecasts with 21 members were also performed for the period December 2017 to February 2018 with the single-layer and the multi-layer snow scheme. The ensemble forecasts are initialized every day at 00UTC using the same procedure described for the deterministic forecasts. The horizontal resolution is about 30km (TCo399) and 91 vertical levels are used. 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 (~10m) and the surface according to:

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

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

The transfer coefficients, C_M and C_H, used to compute the surface stress, τ, and the SHF, are based on Monin Obukhov (M-O) Similarity theory, are a function of the roughness length of momentum and heat, z_{oM} and z_{oH}, and the bulk richardson number, Ri_b, based on Beljaars and Holtslag (1991).

The atmosphere 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 is seen by the atmosphere. The ground heat flux (between the atmosphere and the snow, GHF) in the surface energy balance is calculated for each tile according to:

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

where T_{sfc} is the surface temperature for each tile 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 varies by tile. The agregated value of GHF, across the two snow tiles, is passed to the snow model, to evolve the snow thermodynamics and melt.

The 2m 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. 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%)

The current snow scheme used in operational forecasts at ECMWF and included in HTESSEL is an energy balance model describing the temporal evolution of the heat and mass contents of the snowpack. 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 parametrizations 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 parametrisation by Douville et al. (1995). The gridbox snow cover fraction is parametrized 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 parametrizations 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 multi-layer 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 5 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, Fig 1). Liquid water content is also computed prognosticaly in the multi-layer model, compared to the previous scheme where it was computed diagnostically based on snow temperature.

In addition to the multi-layer formulation several additional parameterisations are included in the new model. (I) The heat conductivity is parametrized 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 parametrized 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 parametrized 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 80cm, 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 ideal 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:

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

where ϵ(=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 use 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 20cm (CHF) and the heat flux convergence (HFC) in the top 20cm of snow. This CHF is calculated according to:

$$CHF = -k_{eff} \frac{\partial T}{\partial z} \quad (6)$$

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 (1997). The HFC is calculated according to:

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

Where c_{ice} is the specific heat capacity of ice, ρ is the average density of the top 20cm 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. Note that equivalent methods exist to calculate the GHF for snow-free soil (e.g. Liebethal and Foken, 2007).

The winter 2013-14 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.

2.3 Process-Oriented diagnostics

Miller et al. (2018) recently proposed a new set of PODs based on the idea that surface energy budget can be split into a ‘driving term’ and ‘response terms’. The idea is that the energy budget can be divided into a driving term: $LW\!\downarrow + SW_{net}$, which varies with synoptic situation, and response terms: *SHF, LHF, GHF and* $-LW\!\uparrow$. Using $LW\!\downarrow + SW_{net}$ instead of the total net radiation removes the explicit dependence on the surface temperature (through $LW\!\uparrow$) from the driving term. The relationship between the driving term and each response term can be summarised with regression coefficients, e.g. for the *SHF*:

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

where each of the α’s can be interpreted as a coupling strength parameter between the driving term and each response term. By substituting the right-hand side of these equations into equation 1 one can derive the following expression relating the α’s:

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

where ϵ is the sum of the β terms divided by the driving term. 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. $¿\alpha_{GHF_{mod}} + \alpha_{SHF_{mod}} + \alpha_{LHF_{mod}} \vee ¿ < ¿\alpha_{GHF_{obs}} + \alpha_{SHF_{obs}} + \alpha_{LHF_{obs}} \vee ¿$) then $¿\alpha_{-LW\uparrow} \vee ¿$, i.e. 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 ($\alpha_{GHF_{mod}\vee\text{¿¿}}$) 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 behaviour of the SEB can be diagnosed in isolation without the confounding effects of other sources of errors, 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:

$$z=\frac{\alpha_{mod}-\alpha_{obs}}{S_{\alpha_{mod}-\alpha_{obs}}}, \quad (10)$$

where $S_{\alpha_{mod}-\alpha_{obs}}=\sqrt{S_{\alpha_{mod}}^{2}+S_{\alpha_{obs}}^{2}}, S_{\alpha}^{2}=\frac{1}{n-2}\frac{\sum(y-y')^{2}}{\sum(x-\bar{x})^{2}}$, 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.

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

3 Results

3.1 Evaluation against conventional weather stations

An anticipated outcome using the multi-layer instead of the single-layer snow scheme is a reduction in the mean error of 2m temperature forecasts over snow-covered surfaces. An evaluation of the change in 2m-temperature forecast skill between the two model formulations against SYNOP stations is performed over the Arctic region (above 65N). 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 Fig 12 of Arduini et al., 2019). The fraction of gridcells in mid-latitudes with values of the CRPS>5K for 2m temperature at a lead time of 5 days is one of ECMWF’s headline scores. 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 (Fig 2a, Fig 3a and 4a), 2m-temperature forecasts with the SL model exhibit a warm bias of 1.7C at both Sodankylä and Summit, with the bias being largest for coldest temperatures. Atlaskin and Vihma (2012) present a multi-centre analysis for eastern Scandinavia 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 2m-temperature observed at various locations across the station, including open and forested sites, show very similar error characteristics (Fig S2).

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

3.2.2 Partitioning sources of 2m-temperature error

As $LW\downarrow + SW_{net}$ is a major driver of 2m-temperature, errors in 2m-temperature are either due to errors in the driving term itself, the relationship between $LW\downarrow + SW_{net}$ and 2m-temperature, or a combination of both (assuming that errors in advection are negligible). Mean errors in the radiative forcing term are positive at Sodankylä (~6Wm^{-2}), particularly for low values of this term, and therefore contribute to the positive temperature errors (see Fig 3b). The mean error in the radiation term is negative at Summit (~8Wm^{-2}), shows that radiation errors are not responsible for the positive mean temperature bias there (see Fig 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 (Fig S3).

At both sites the 2m-temperature in the SL forecasts is less sensitive to changes in $LW\downarrow + SW_{net}$ than it is in observations (0.13K/Wm^{-2} compared to 0.17K/Wm^{-2} at Sodankylä and 0.14K/Wm^{-2} compared to 0.19K/Wm^{-2} at Summit). The inclusion of the multi-layer snow increases the sensitivity of 2m-temperature to radiative forcing at both sites. The lack of any substantial change in the driving term (Figures 3e & 4e cf. 3b & 4b) suggests that the reduction in error is due to an improvement in the response of 2m-temperature to radiative forcing. At low values of the $LW\downarrow + SW_{net}$ the values of 2m-temperature are lower for the ML experiment, which goes hand in hand with improved forecasts of cold conditions. The sensitivity at Summit is slightly too high in the ML experiment and still too low at Sodankylä.

3.2.3 Surface energy budget process relationships

The responsiveness of 2m-temperature to the radiative forcing is closely related to the responsiveness of the surface temperature. Indeed, the surface-temperature-$LW{\downarrow}+SW_{net}$ diagrams closely resemble those for 2m-temperature. Surface temperature is too insensitive to variations in the radiative forcing in the SL forecasts at both sites (Fig 5a & Fig 6a). This sensitivity increases at both sites in the ML forecasts but remains too low at Sodankylä (Fig 5d) and becomes too high at Summit (Fig 6d).

Because the energy budget is closed, an under or over-responsive surface temperature (or $LW{\uparrow}$ equivalently) to radiative forcing must be due to the remaining response terms (SHF, LHF or GHF) over or under-responding respectively. By looking at the response of these fluxes to variations in $LW{\downarrow}+SW_{net}$ we can understand the causes of systematic errors in the surface temperature sensitivity, and how this changes between model versions, from a process perspective.

To help 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 (Figs S4 and S5). The surface-temperature sensitivity to radiative forcing is higher in non-turbulent regimes (Ri>0.25) than in turbulent regimes (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. As Ri increases, the turbulent fluxes decrease and hence the fraction of incoming radiation they can balance (i.e. $¿\alpha_{SHF}+\alpha_{LHF}\vee¿$) decreases. The fraction balanced by LW↑ and GHF ($¿\alpha_{GHF}+\alpha_{-LW\uparrow}\vee¿$) must therefore increase, allowing the surface temperature to become more responsive. This implies that a model with excessive turbulent transport, for example, would have a surface-temperature sensitivity that was too low.

In the SL forecast the coupling strength to the land-surface is too strong at both sites (i.e. the fraction of the radiative forcing going into heating the land surface is too large): $¿\alpha_{GHF_{mod}}\vee¿\vee\alpha_{GHF_{obs}}\vee¿$, see Figs 5c and 6c). The coupling to the atmosphere is also too high at Sodankyla (i.e. $¿\alpha_{SHF_{mod}}+\alpha_{LHF_{mod}}\vee¿\vee\alpha_{SHF_{obs}}+\alpha_{LHF_{obs}}\vee¿$, see Fig 5b , S6 and Table 1 and 2), which results in the surface temperature sensitivity being too low (i.e. $¿\alpha_{-LW\uparrow_{mod}}\vee¿\vee\alpha_{-LW\uparrow_{obs}}\vee¿$). At Summit the coupling to the atmosphere is too low (and $\alpha_{SHF_{mod}}$ even has the wrong sign, see Fig 6b) but because $¿\alpha_{SHF_{mod}}+\alpha_{LHF_{mod}}+\alpha_{GHF_{mod}}\vee¿$ is too high overall (See Fig 6b, S7 and Table 2), the surface-temperature response is also too low, as it is at Sodankylä.

Using the multi-layer instead of the single-layer snow scheme directly influences the coupling between the radiation and the GHF, i.e. α_{GHF}, because the snow temperature used in the GHF calculation (Eq 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 $¿\alpha_{GHF_{mod}}\vee¿$) at both sites (see Fig 5 & 6 and Tables 1 & 2). As a result, this leads to an increased and improved surface-temperature sensitivity at both sites. However, $¿\alpha_{GHF_{mod}}\vee¿$ remains a bit too high at Sodankyla while it becomes too low at

Summit. The reduction in the magnitude of $\alpha_{GHF_{mod}}$ is also much larger at Summit than at Sodankyla. This 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 Eq 4), for snow under forest is about three-times that for exposed snow (20 Wm^{-2} compared to 7). As a result, a larger GHF will be maintained over the forested tile, compared to a 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 parametrizations 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 (Fig 6b and Table 2). Coupling to the multi-layer snow changes the sign, bringing $\alpha_{SHF_{mod}}$ into close agreement with the observed value (Fig 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 (Fig 7b & d) which subsequently improves the distribution of SHF (Fig 7a & c) and its response to variations in radiative forcing.

The ability of a change in one of the model's parametrizations (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.

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 is too responsive too much to changes in radiative forcing, and as a result T_{sfc} still does not respond to radiative forcing as much as in observations, even with the ML snow.

4. Discussion: the role of 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 $¿\alpha_{SHF_{mod}}+\alpha_{LHF_{mod}}\vee ¿\vee\alpha_{SHF_{obs}}+\alpha_{LHF_{obs}}\vee ¿$). This could either be because of errors in the formulation of the turbulent exchange in the surface layer (between 10m and the surface) or in the outer layer (i.e. above 10m). 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 Eq 2 & 3, given both the observed flux and bulk properties at a given site (for example see Tjernström et al., 2005). In practice however, in vegetated areas or complex terrains, 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 (Fig 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 multi-layer snow scheme. This reduces the fraction of radiation being balanced by the SHF, $|\alpha_{SHF} \vee$ ¿, 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 Europ (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 Eq 3).

5. Conclusions

In this study we have presented a new way to evaluate model developments from the perspective of SEB process-relationships for surface & 2m 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, in winter: Summit station, in the centre of the Greenland Ice Sheet and Sodankylä, a heterogeneous Arctic Taiga site in Finland. However, 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 2m-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 2m 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 analysing 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 multi-layer snow scheme results in an overall improvement in Arctic 2m-temperature forecasts, reducing a systematic warm bias, particularly during cold events. Improvements in the mean 2m-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 (Eq 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 behaviour 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, 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: Fig 1), simplifying the interpretation of the analysis. However, these diagnostics could be usefully applied to mid-latitudes, 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).

Acknowledgments

The work described in this article has received funding fromthe European Union's Horizon 2020 Research and Innovation programme through grant agreement No. 727862 APPLICATE. The content of the article is the sole responsibilityof the author(s) and it does not represent the opinion of theEuropean 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.

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. doi:10.18739/A2Z37J]. The Sodankyla surface energy budget data is 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 multi-layer snow model will be published at the Zenodo repository, following acceptance of the manuscript, with the following doi:10.5281/zenodo.3755373

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).

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, 2019MS001725. 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

Bauer, P., Magnusson, L., Thépaut, J.−N. and Hamill, T.M. (2016), Aspects of ECMWF model performance in polar areas. Q.J.R. Meteorol. Soc., 142: 583-596. doi:10.1002/qj.2449

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: Verification922from field site to terrestrial water storage and impact in the integrated forecast system. Journal of hydrometeorology,10(3), 623–643.

Beljaars, A., Balsamo, G., Bechtold, P., Bozzo, A., Forbes, R., Hogan, R.J., Köhler, M., Morcrette, J.-J., Tompkins, A.M., Viterbo, P., Wedi, N., 2018. The Numerics of Physical

Parametrization in the ECMWF Model. Front. Earth Sci. 6. https://doi.org/10.3389/feart.2018.00137

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.

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

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.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge−Sanz, B.M., Morcrette, J.−J., Park, B.−K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.−N. and Vitart, F. (2011), The ERA−Interim reanalysis: configuration and performance of the data assimilation system. Q.J.R. Meteorol. Soc., 137: 553-597. doi: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),972899–916.

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.

Essery, R. (2010). Snow parameterisation in GCMs. In R. L. Armstrong & E. Brun (Eds.), Snow and Climate. Cambridge University Press.

Essery, R., Rutter, N., Pomeroy, J., Baxter, R., Stahli, M., Gustafsson, D., et al. (2009). SnowMIP2: An evalution of forest snow process simulation. Bulletin of the American Meteorological Society, 90(8), 1130–1135. https://doi.org/10.1175/2009BAMS2629.1

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

Geer, A. J. (2016). Significance of changes in medium-range forecast scores. Tellus, Series A: Dynamic Meteorology and Oceanography, 68(1), 30229. https://doi.org/10.3402/tellusa.v68.30229

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., R.J. Hogan, E. O'Connor, D. Bouniol, M.E. Brooks, J. Delanoé, D.P. Donovan, J.D. Eastment, N. Gaussiat, J.W. Goddard, M. Haeffelin, H.K. Baltink, O.A. Krasnov, J. Pelon, J. Piriou, A. Protat, H.W. Russchenberg, A. Seifert, A.M. Tompkins, G. van Zadelhoff, F. Vinit, U. Willén, D.R. Wilson, and C.L. Wrench, 2007: Cloudnet. Bull. Amer. Meteor. Soc., 88, 883–898, https://doi.org/10.1175/BAMS-88-6-883

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., Dawson, J., Doblas-Reyes, F., Fairall, C., Goessling, H.F., Holland, M., Inoue, J., Iversen, T., Klebe, S., Lemke, P., Losch, M., Makshtas, A., Mills, B., Nurmi, P., Perovich, D., Reid, P., Renfrew, I.A., Smith, G., Svensson, G., Tolstykh, M., Yang, Q., 2016. Advancing Polar Prediction Capabilities on Daily to Seasonal Time Scales. Bull. Am. Meteorol. Soc. 97, 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

Krinner, G., Derksen, C., Essery, R., Flanner, M., Hagemann, S., Clark, M., et al. (2018). ESM-SnowMIP: Assessing snow models and quantifying snow-related climate feedbacks. Geoscientific Model Development. https://doi.org/10.5194/gmd-11-5027-2018

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. J. Geophys. Res. Atmospheres 124, 4975–4984. https://doi.org/10.1029/2019JD030454

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

Liebethal, C., Foken, T., 2007. Evaluation of six parameterization approaches for the ground heat flux. Theor. Appl. Climatol. Wien 88, 43–56. http://dx.doi.org/10.1007/s00704-005-0234-0

Lynch-Stieglitz, M. (1994). The development and validation of a simple snow model for the GISS GCM. Journal of Climate, 7(12), 1842–1855.

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. 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(10), 4777–4796. https://doi.org/10.1029/2017JD027377

Oke, T. R. (1987). Boundary layer climates, Second edition. 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(7), 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 C: Oceans, 107(10), 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(2), 117–133. https://doi.org/10.1002/jame.20013

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. 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(3), 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: Atmospheres*, *104*(D16), 19,587–19,597.

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. and Teixeira, J. (1999), The representation of soil moisture freezing and its impact on the stable boundary layer. Q.J.R. Meteorol. Soc., 125: 2401-2426. doi:10.1002/qj.49712555904

Working group on numerical experimentation (WGNE): Systematic Error Survey Results Summary (Feb 2019): https://www.wcrp-climate.org/JSC40/12.7b.%20WGNE_Systematic_Error_Survey_Results_20190211.pdf

Zadra, A., K. Williams, A. Frassoni, M. Rixen, Á.F. Adames, J. Berner, F. Bouyssel, B. Casati, H. Christensen, M.B. Ek, G. Flato, Y. Huang, F. Judt, H. Lin, E. Maloney, W. Merryfield, A. Van Niekerk, T. Rackow, K. Saito, N. Wedi, and P. Yadav, 2018: Systematic Errors in Weather and Climate Models: Nature, Origins, and Ways Forward. Bull. Amer. Meteor. Soc., 99, ES67–ES70, https://doi.org/10.1175/BAMS-D-17-0287.1

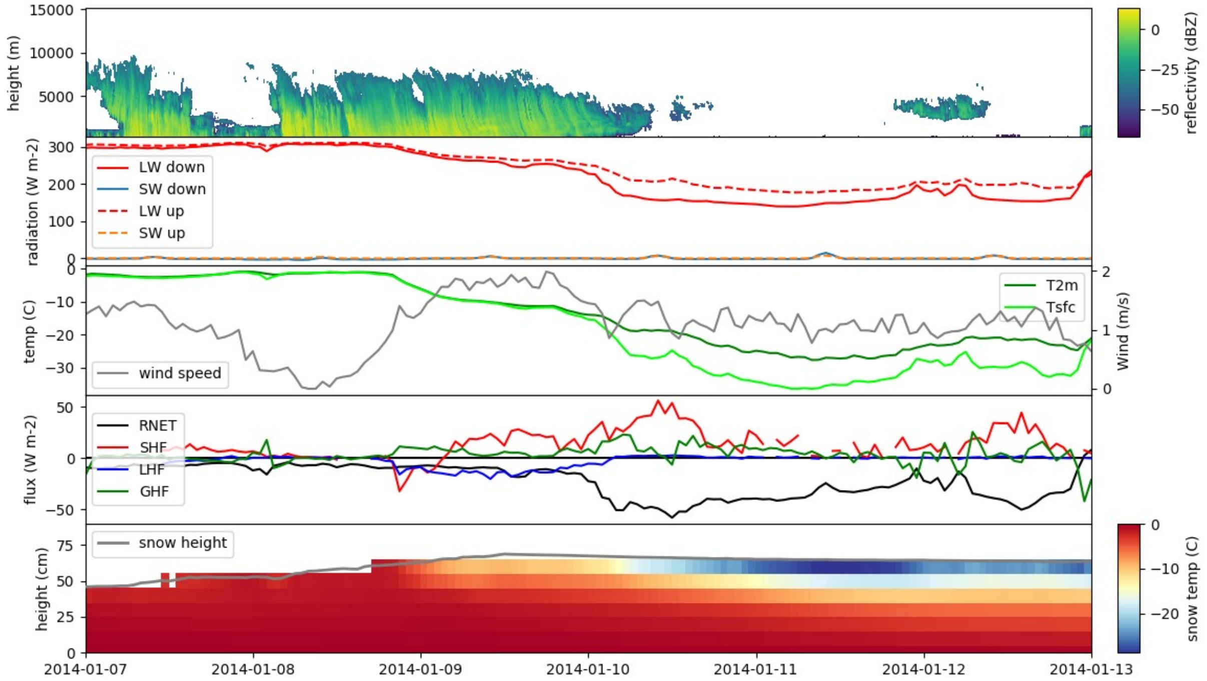

Figure 1: Observed meteogram for a case study at Sodankylä, Finland, in Jan 2014. It shows (from top-to-bottom) cloud radar reflectivity (from CloudNet; Illingworth et al. (2007)); radiation terms; wind speed, surface and 2m temperature; energy balance terms: total net radiation (RNET), sensible (SHF), 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).

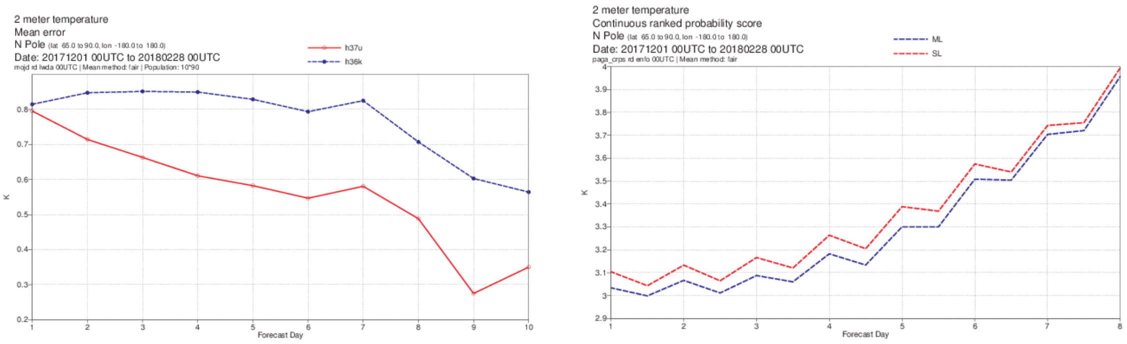

Figure 2. 00UTC 2m temperature mean error for 10-day deterministic forecasts for DJF 2013/14 (left) and 2m-temperature Continuous Ranked Probability Score for 8-day ensemble forecasts for DJF 2017/18 (CRPS; right) for the Arctic region (>65N), compared to SYNOP. Forecasts with single layer snow are shown in blue and multi-layer snow are shown in red.

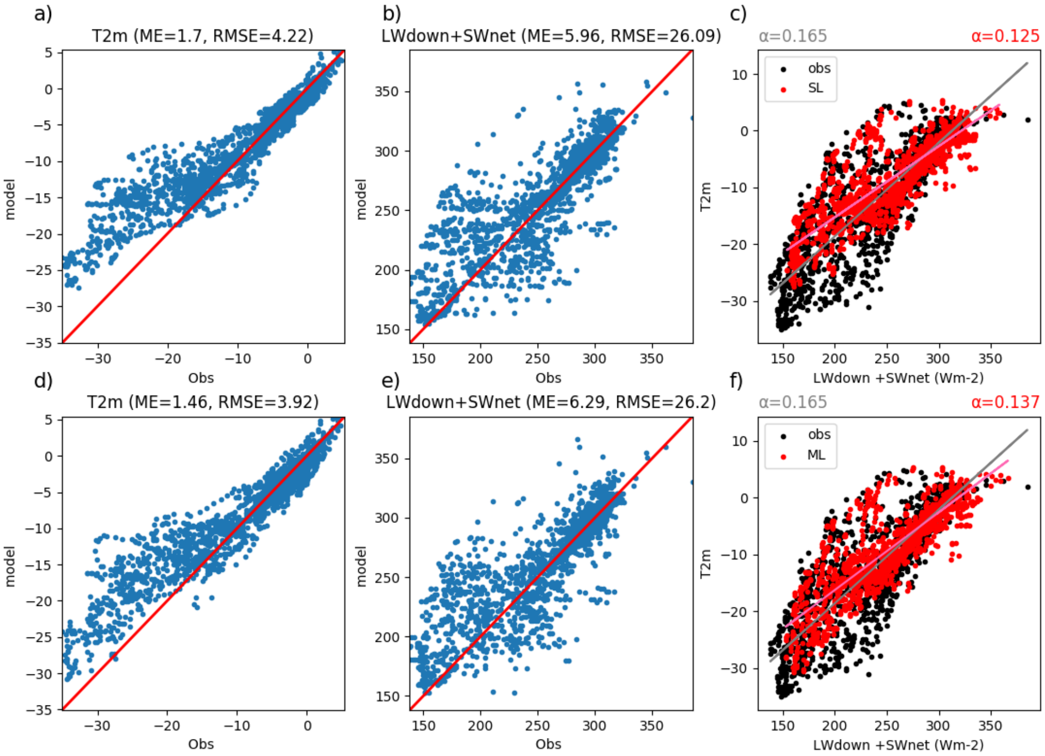

Figure 3. Hourly observed vs forecast (during day-2) 2m temperature (a & d), $LW\downarrow+SW_{net}$ (b & e), and the relationship between them (c & f) in observations (black) and each model formulation (red) for Sodankylä with single layer snow (top row) and multi-layer snow (bottom row) for DJF 2013/14. The regression coefficient is shown for the observations (black text) and the models (red text).

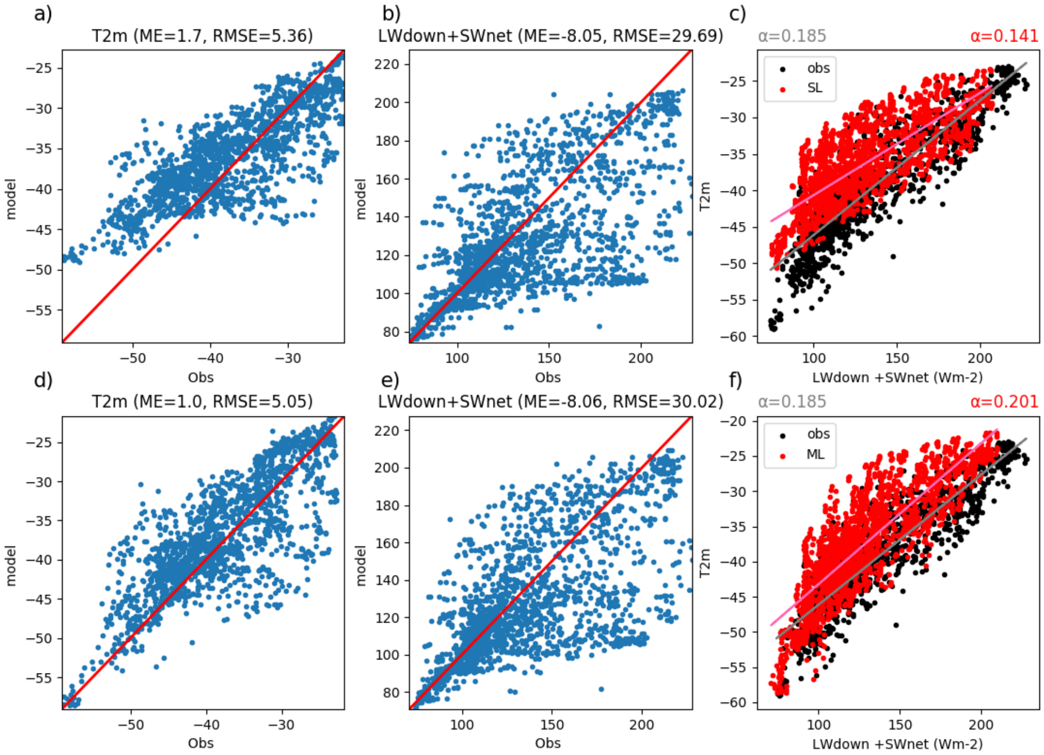

Figure 4. Hourly observed vs forecast (during day-2) 2m temperature (a & d), LW↓+SW$_{net}$ (b & e), and the relationship between them (c & f) in observations (black) and each model formulation (red) for Summit with single layer snow (top row) and multi-layer snow (bottom row) for DJF 2013/14. The regression coefficient is shown for the observations (black text) and the models (red text).

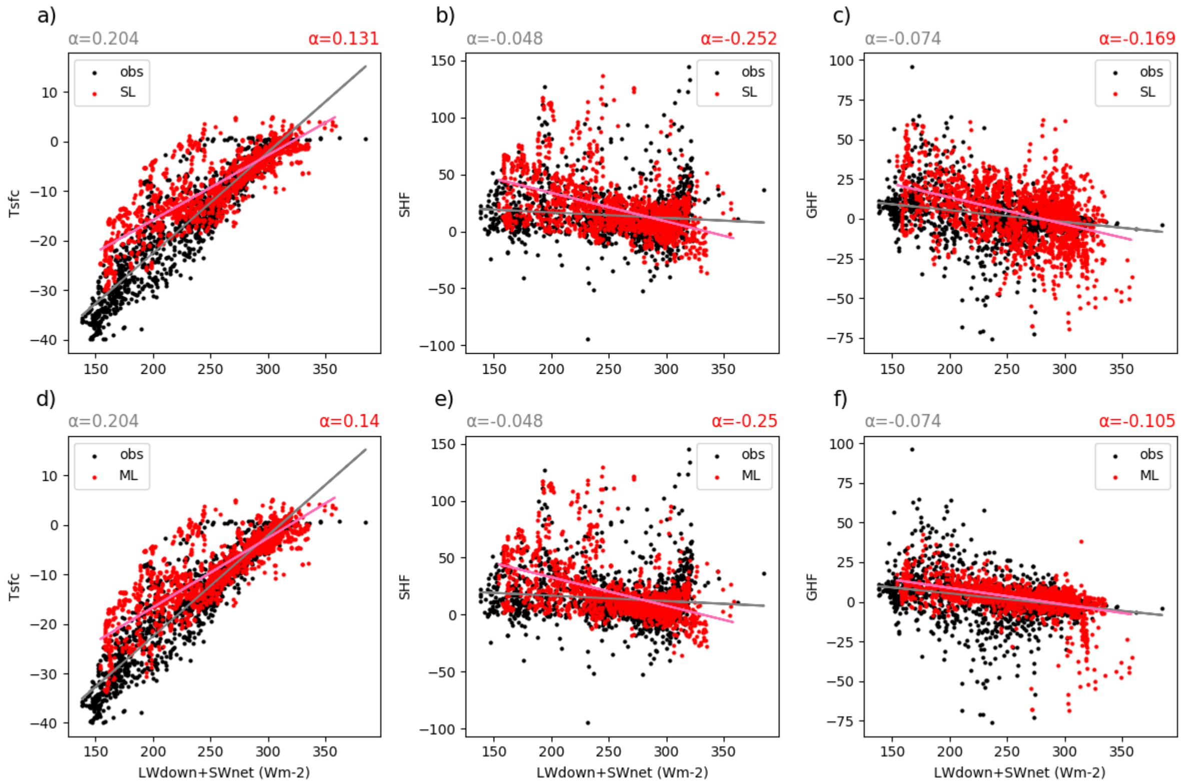

Figure 5. process relationship diagrams and sensitivity parameters for surface temperature (T_{sfc}; left), sensible heat flux (SHF; middle) and ground heat flux (GHF; right) for Sodankyla, Finland. Observed values are shown in black, model values are shown in red for single layer snow (a-c) and multi-layer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

	regression parameter (z-statistic)		
Parameter	**Observations**	**SL**	**ML**
T_{sfc}	0.20	0.131 (z=-34.3, p=0.00)	0.140 (z=**-29.0**, p=0.00)
SHF	-0.048	-0.252 (z=**-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=**-4.99**, p=2.97e-7)
LHF	-0.053	-0.028 (z=4.97, p=3.41e-7)	-0.033 (z=**4.39**, p=5.51e-6)
-LW↑	-0.79	-0.55 (z=-29.0, p=0.00)	-0.58 (z=**-24.7**, p=0.00)
T2m	0.165	0.125 (z=-16.3, p=0.00)	0.133 (z=**-12.8**, p=0.00)

Table 1. Observed and modelled regression parameters at Sodankylä. Bold values highlight which z-score is better.

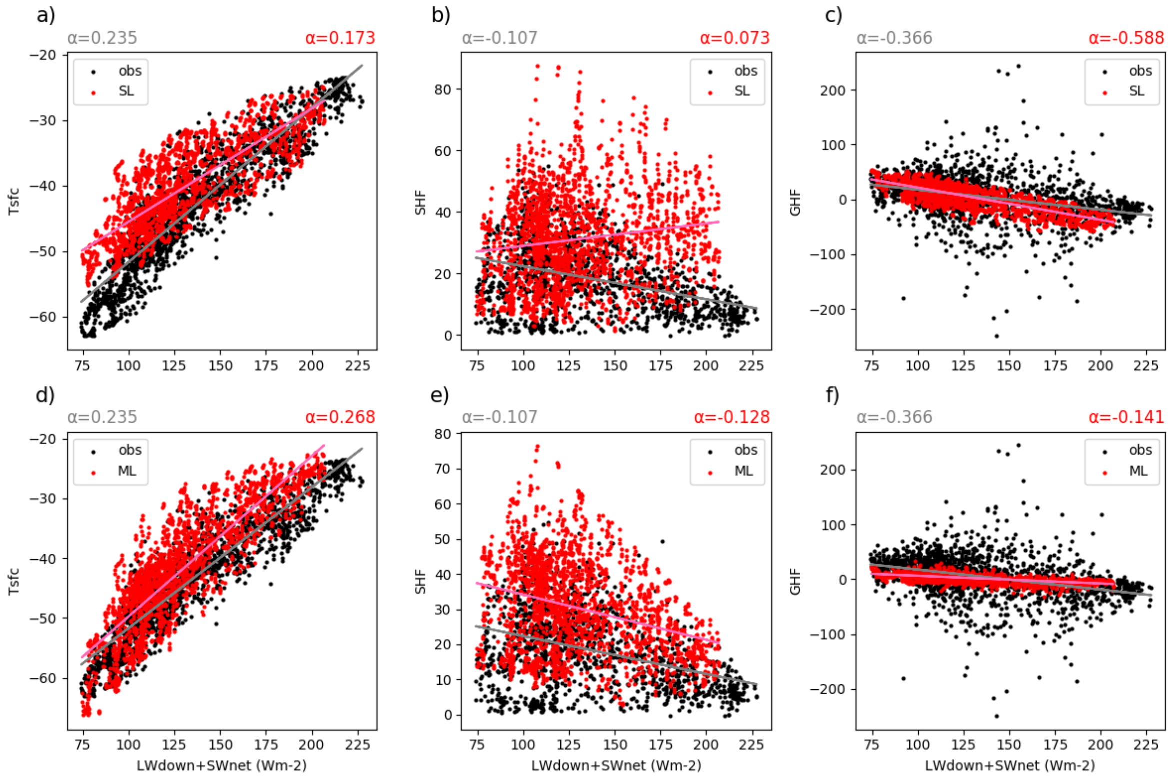

Figure 6. process relationship diagrams and sensitivity parameters for surface temperature (T_{sfc}; left), sensible heat flux (SHF; middle) and ground heat flux (GHF; right) for Summit, Greenland. Observed values are shown in black, model values are shown in red for single layer snow (a-c) and multi-layer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

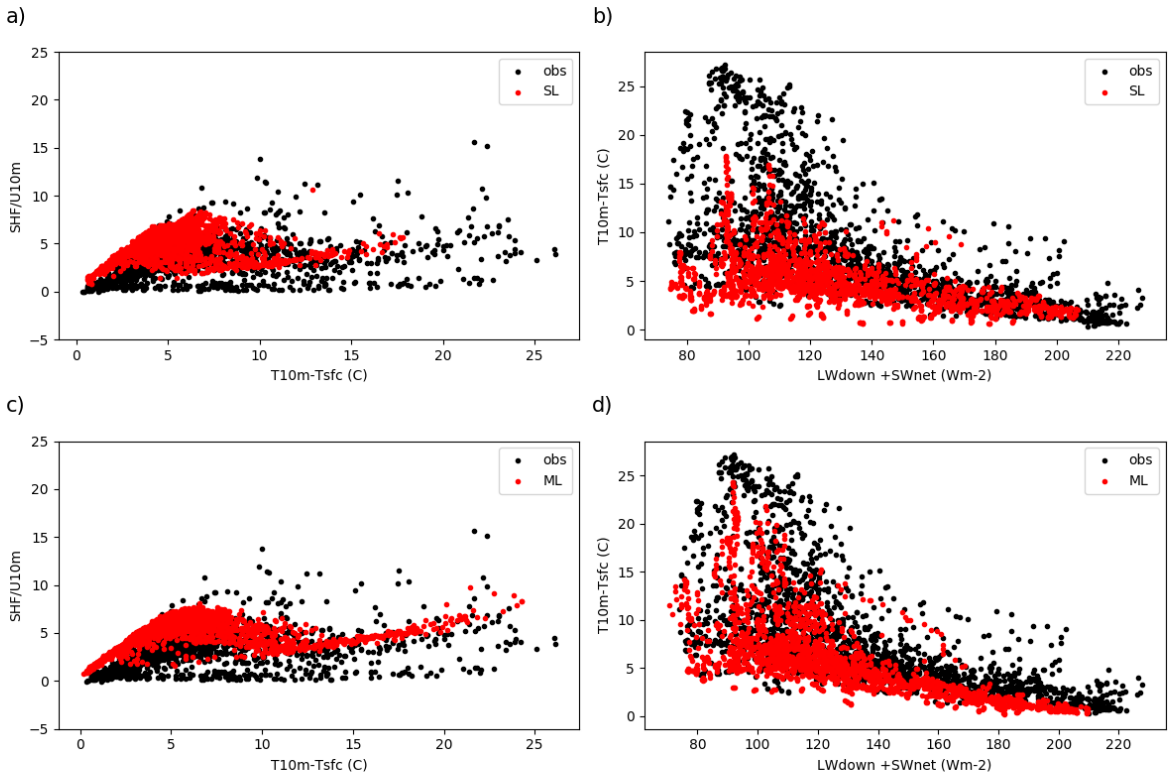

Figure 7. Sensible heat, scaled by wind speed, as a function of inversion strength at Summit from forecasts with the single-lager model (SL, a) and multi-layer model (ML, c). Inversion strength as a function of radiative forcing (LW↓ + SW$_{\text{net:}}$) for SL (b) and ML (d). Observations are shown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