Scientific and human errors in a snow model intercomparison

Scientific and human errors in a snow model intercomparison

Cecile B. Menard[1], Richard Essery[1], Gerhard Krinner[2], Gabriele Arduini[3], Paul Bartlett[4], Aaron Boone[5], Claire Brutel-Vuilmet[2], Eleanor Burke[6], Matthias Cuntz[7], Yongjiu Dai[8], Bertrand Decharme[4], Emanuel Dutra[9], Xing Fang[10], Charles Fierz[11], Yeugeniy Gusev[12], Stefan Hagemann[13], Vanessa Haverd[14], Hyungjun Kim[15], Matthieu Lafaysse[16], Thomas Marke[17], Olga Nasonova[12], Tomoko Nitta[15], Masashi Niwano[18], John Pomeroy[10], Gerd Schädler[19], Vladimir Semenov[20], Tatiana Smirnova[21], Ulrich Strasser[17], Sean Swenson[22], Dmitry Turkov[23], Nander Wever[24,11], Hua Yuan[8]

[1]School of Geosciences, University of Edinburgh, Edinburgh, United Kingdom

[2]CNRS, Université Grenoble Alpes, Institut de Géosciences de l'Environnement (IGE), Grenoble, France

[3]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, United Kingdom

[4]Climate Research Division, Environment and Climate Change Canada, Toronto, Canada

[5]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

[6]Met Office Hadley Centre, FitzRoy Road, Exeter, United Kingdom

[7]Université de Lorraine, AgroParisTech, INRAE, UMR Silva, Nancy, France

[8]School of Atmospheric Sciences, Sun Yat-sen University, Guangzhou, China

[9]Instituto Dom Luiz, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

[10]Centre for Hydrology, University of Saskatchewan, Saskatoon, Canada

[11]WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland

[12]Institute of Water Problems, Russian Academy of Sciences, Moscow, Russia

[13]Institut für Küstenforschung, Helmholtz-Zentrum Geesthacht, Geesthacht, Germany

Early Online Release: This preliminary version has been accepted for publication in *Bulletin of the American Meteorological Society*, may be fully cited, and has been assigned DOI 10.1175/BAMS-D-19-0329.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

[14]CSIRO Oceans and Atmosphere, Canberra, ACT, Australia

[15]Institute of Industrial Science, the University of Tokyo, Tokyo, Japan

[16]Grenoble Alpes, Université de Toulouse, Météo-France, CNRS, CNRM, Centre d'Etudes de la Neige, Grenoble, France

[17] Department of Geography, University of Innsbruck, Innsbruck, Austria

[18]Meteorological Research Institute, Japan Meteorological Agency, Tsukuba, Japan

[19]Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany

[20]A.M. Obukhov Institute of Atmospheric Physics, Russian Academy of Sciences, Moscow, Russia

[21]Cooperative Institute for Research in Environmental Science/Earth System Research Laboratory, NOAA, Boulder, CO, USA

[22]Advanced Study Program, National Center for Atmospheric Research, Boulder, Colorado

[23]Institute of Geography, Russian Academy of Sciences, Moscow, Russia

[24]Department of Atmospheric and Oceanic Sciences, University of Colorado Boulder, Boulder, CO, USA

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

Abstract

Twenty-seven models participated in the Earth System Model - Snow Model Intercomparison Project (ESM-SnowMIP), the most data-rich MIP dedicated to snow modelling. Our findings do not support the hypothesis advanced by previous snow MIPs: evaluating models against more variables, and providing evaluation datasets extended temporally and spatially does not facilitate identification of key new processes requiring improvement to model snow mass and energy budgets, even at point scales. In fact, the same modelling issues identified by previous snow MIPs arose: albedo is a major source of uncertainty, surface exchange parametrizations are problematic and individual model performance is inconsistent. This lack of progress is attributed partly to the large number of human errors that led to anomalous model behaviour and to numerous resubmissions. It is unclear how widespread such errors are in our field and others; dedicated time and resources will be needed to tackle this issue to prevent highly sophisticated models and their research outputs from being vulnerable because of avoidable human mistakes. The design of and the data available to successive snow MIPs were also questioned. Evaluation of models against bulk snow properties was found to be sufficient for some but inappropriate for more complex snow models whose skills at simulating internal snow properties remained untested. Discussions between the authors of this paper on the purpose of MIPs revealed varied, and sometimes contradictory, motivations behind their participation. These findings started a collaborative effort to adapt future snow MIPs to respond to the diverse needs of the community.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

Capsule

The latest snow model intercomparison identified the same modelling issues as previous iterations over 23 years. Lack of new insights are attributed partly to human errors and intercomparison projects design.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

1. Introduction

The Earth System Model-Snow Model Intercomparison Project (ESM-SnowMIP; Krinner et al., 2018) is the third in a series of MIPs spanning seventeen years investigating the performance of snow models. It is closely aligned with the Land Surface, Snow and Soil Moisture Model Intercomparison Project (LS3MIP; van den Hurk et al. 2016), which is a contribution to the sixth Coupled Model Intercomparison Project (CMIP6). The Tier 1 reference site simulations (Ref-Site in Krinner et al., 2018), the results of which are discussed in this paper, is the first of ten planned ESM-SnowMIP experiments and the latest iteration of MIPs using in situ data for snow model evaluation. The Project for Intercomparison of Land surface Parameterization Schemes Phase 2(d) (PILPS 2(d)) was the first comprehensive intercomparison focusing on the representation of snow in land surface schemes (Pitman and Henderson-Sellers, 1998; Slater et al., 2001) and evaluated models at one open site for 18 years. It was followed by the first SnowMIP (hereafter SnowMIP1; Etchevers et al., 2002; Etchevers et al., 2004), which evaluated models at four open sites for a total of 19 site-years and by SnowMIP2 (Rutter et al., 2009; Essery et al., 2009) which investigated simulations at five open and forested site pairs for 9 site-years.

Twenty-seven models from twenty-two modelling teams participated in the ESM-SnowMIP Ref-Site experiment (ESM-SnowMIP hereafter). A short history of critical findings in previous MIPs is necessary to contextualise the results. PILPS 2(d) identified sources of model scatter to be albedo and fractional snow cover parametrizations controlling the energy available for melt, and longwave radiative feedbacks controlled by exchange coefficients for sensible and latent heat fluxes in stable conditions (Slater et al., 2001). SnowMIP1

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

corroborated the latter finding, adding that the more complex models were better able to simulate net longwave radiation but both complex models and simple models with appropriate parametrizations were able to simulate albedo well (Etchevers et al, 2004) (Baartman et al., 2020, showed that there is no general consensus about what "model complexity" is; for clarity, we will define models explicitly incorporating larger numbers of processes, interactions and feedbacks as more complex). SnowMIP2 found little consistency in model performance between years or sites and, as a result, there was no subset of better models (Rutter et al., 2011). The largest errors in mass and energy balances were attributed to uncertainties in site-specific parameter selection rather than to model structure. All these projects concluded that more temporal and spatial data would improve our understanding of snow models and reduce the uncertainty associated with process representations and feedbacks on the climate.

This paper discusses results from model simulations at five mountain sites (Col de Porte, France; Reynolds Mountain East, Idaho, USA; Senator Beck and Swamp Angel, Colorado, USA; Weissfluhjoch, Switzerland), one urban maritime site (Sapporo, Japan) and one Arctic site (Sodankylä, Finland); results for three forested sites will be discussed in a separate publication. Details of the sites, forcing and evaluation data are presented in Menard et al. (2019). Although the 97 site-years of data for these seven reference sites may still be insufficient, they do respond to the demands of previous MIPs by providing more sites in different snowy environments over more years.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

2. The false hypothesis

In fiction, a false protagonist is one who is presented as the main character but turns out not to be, often by being killed off early (e.g. Marion Crane in Psycho, 1960; Dallas in Alien, 1979; Ned Stark in A Game of Thrones, Martin, 1996). This narrative technique is not used in scientific literature, even though many scientific hypotheses advanced in project proposals are killed early at the research stage. Most scientific journals impose strict manuscript composition guidelines to encourage research studies to be presented in a linear and cohesive manner. As a consequence, many “killed” hypotheses are never presented, and neither are the intermediary steps that lead to the final hypothesis. This is an artifice that we all comply with even though hypothesizing after the results are known (known as HARKing; Kerr, 1998) is a practice associated with the reproduction crisis (Munafò et al., 2017).

Our working hypothesis was formed at the design stage of ESM-SnowMIP and is explicit in Krinner et al. (2018): more sites over more years will help us to identify crucial processes and characteristics that need to be improved as well as previously unrecognized weaknesses in snow models. However, months of analysing results led us to conclude the unexpected: more sites, more years and more variables do not provide more insight into key snow processes. Instead, this leads to the same conclusions as previous MIPs: albedo is still a major source of uncertainty, surface exchange parametrizations are still problematic, and individual model performance is inconsistent. In fact, models are less classifiable with results from more sites, years and evaluation variables. Our initial, or false, hypothesis had to be killed off.

Developments *have* been made, particularly in terms of the complexity of snow process representations, and conclusions from PILPS2(d) and snow MIPs have undoubtedly driven model development. Table 1 shows that few participating models now have a fixed snow

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

density or thermal conductivity, only two models still parametrize snow albedo as a simple function of temperature, no model uses constant surface exchange coefficients, more models can now represent liquid water in snow, and only three still have a composite snow/soil layer. These changes demonstrate progress for individual models, but they do not for snow science: most of these parametrizations have existed for decades. Differences between models remain, but the range of model complexity is smaller than it was in previous MIPs.

The pace of advances in snow modelling and other fields in climate research is limited by the time it takes to collect long-term datasets and to develop methods for measuring complex processes. Furthermore, the logistical challenges of collecting reliable data in environments where unattended instruments are prone to failure continue to restrict the spatial coverage of quality snow datasets.

False protagonists allow narrators to change the focus of the story. Our “false hypothesis” allows us to re-focus our paper not on what the model results are – doing so would merely repeat what previous snow MIPs have concluded – but on why, in the twenty four years since the start of PILPS 2 (d), the same modelling issues have repeatedly limited progress in our field, when other fields relying on technology and computing have changed beyond recognition.

3. <u>The Beauty Contest</u>

Ranking models (or the “beauty contest”, as insightfully described by Ann Henderson-Sellers when presenting results from PILPS) offers little or no insight into their performance,

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

but it has become the compulsory starting point for presenting MIP results. Figures 1 and 2 show models ranked according to errors in daily averages of snow water equivalent (SWE), surface temperature, albedo and soil temperature (note that not all of these variables were measured at all sites or output by all models). To avoid errors in simulations for snow-free or partially snow-covered ground, errors in albedo and surface and soil temperatures were only calculated for periods with measured snow depths greater than 0.1 m and air temperatures below 0°C. Measured and modelled snow surface temperatures greater than 0°C and albedos less than 0.5 were excluded from the error calculations. Bias is shown for SWE, surface temperature, albedo and soil temperature. Root mean square error normalised by standard deviation (NRMSE) is presented only for SWE and surface temperature because standard deviations of albedo and soil temperature are small during periods of continuous snow cover.

Discussion of the results in Sections 3.1 to 3.3. will demonstrate why our initial hypothesis was rejected: no patterns emerge, no sweeping statements can be made. The preliminary conclusion presented in Krinner et al. (2018) that "model complexity per se does not explain the spread in performance" still stands. For example, Table 1 shows that RUC is one of the simplest models, but Figures 1 and 2 show that it often has smaller errors than more complex models. This is not to say that model developments are useless: there are large differences between simulations submitted for older and newer versions of a few models. Errors in SWE – the most commonly used variable for evaluation of site simulations – are greatly reduced in HTESSEL-ML, JULES-UKESM/JULES-GL7 and ORCHIDEE-E/ORCHIDEE-MICT compared with HTESSEL, JULES-I and ORCHIDEE-I, and errors in soil temperature are greatly reduced in JSBACH-PF which, unlike its predecessor JSBACH, includes a soil freezing parametrization. There is little or no reduction in errors for other variables between versions.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

Errors in the ESM-SnowMIP driving and evaluation data are not discussed here because they are discussed in Menard et al. (2019): implicit in the following sections is that a model can only be as good as the data driving it and against which it is evaluated.

3.1 Snow water equivalent and surface temperature

Mean SWE and surface temperature NRMSEs in Figure 1 are generally low: below 0.6 for half of the models and 1 or greater for only four models. Biases are also relatively low: less than 2°C in surface temperature and less than 0.2 in normalised SWE for four out of five sites in Figure 2. The sign of the biases in surface temperature are the same for at least four out of five sites for all except four models (JULES-I, ORCHIDEE-E, ORCHIDEE-MICT and SWAP). The six models with the largest negative biases in SWE are among the seven models that do not represent liquid water in snow. The seventh model, RUC, has its largest negative bias at Sapporo, where rain-on-snow events are common. Wind-induced snow redistribution, which no model simulates at a point, is partly responsible for Senator Beck being one of the two sites with largest SWE NRMSE in more than half of the models.

Four of the best models for SWE NRMSE are among the worst for surface temperature NRMSE (SPONSOR, Crocus, CLASS and HTESSEL-ML). Decoupling of the snow surface from the atmosphere under stable conditions is a long-standing issue which Slater et al. (2001) investigated in PILPS 2(d). Underestimating snow surface temperature leads to a colder snowpack that takes longer to melt and remains on the ground for longer. In 2001, most models used Richardson numbers to calculate surface exchange; in 2019, most use Monin-Obukhov similarity theory (MOST). However, assumptions of flat and horizontally

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

homogeneous surfaces and steady-state conditions in MOST make it inappropriate for describing conditions not only over snow surfaces, but also over forest clearings and mountains: in other words, at all sites in this study. Exchange coefficient are commonly used to tune near-surface temperature in numerical weather prediction models even if to the detriment of the representation of stable boundary layers (Sandu et al., 2013). Conway et al. (2018) showed that such tuning in snowpack modelling improved surface temperature simulations but at the expense of overestimating melt. It is beyond the scope of this paper (and in view of the discussion on sources of errors in Section 4, possibly beyond individual modelling teams) to assess how individual models have developed and evaluated their surface exchange and snowpack evolution schemes. However, differences in model ranking between SWE and surface temperature suggest that this issue is widespread and warrants further attention.

3.2 Albedo

Errors in modelled winter albedo (Li et al., 2016) and implications for snow albedo feedback on air temperature (Randall et al., 2007; Flato et al., 2013) have been linked to errors in snow cover fraction (SCF) (e.g. Roesch et al, 2006) and vegetation characteristics in the boreal regions, rather than to the choice or complexity of snow albedo schemes (Essery, 2013; Wang et al, 2016). These should not affect ESM-SnowMIP because vegetation characteristics were provided to participants (all sites discussed here are in clearings or open landscapes) and snow cover during accumulation is expected to be complete. However, eleven models did not impose complete snow cover (Figure 3) such that, again, differences in surface albedo

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

are inextricably linked to differences in snow cover fraction; implications are discussed in Section 4.1.

As in previous studies (e.g. Etchevers et al., 2004; Essery, 2013), the specific albedo scheme or its complexity does not determine model performance in ESM-SnowMIP. Neither of the two models with the smallest range of biases, CLASS and EC-Earth, imposed SCF = 1 and both use simple albedo schemes in which snow albedo decreases depending on time and temperature. Snow albedo parametrizations (Table 1) determine rates at which albedo varies, but ranges within which the schemes operate are still determined by user-defined minimum and maximum snow albedos to which models are very sensitive. For most models these parameters are the same at all sites, but measurements suggest that they differ; it is unclear whether some of these variations are due to site-specific measurement errors (e.g. instruments or vegetation in the radiometer field of view). This issue should be investigated further as this is not the first time that model results have been inconclusive because of such uncertainties (e.g. Essery et al., 2013).

3.3 Soil temperature

Five models systematically underestimate soil temperatures under snow (JSBACH MATSIRO, ORCHIDEE-I, RUC and SURFEX-ISBA) and four systematically overestimate them (CLM5, CoLM, JULES-GL7 and ORCHIDEE-MICT), although negative biases are often larger than positive ones. Soil temperatures are not consistently over- or underestimated by all models at any particular site. Three of the models (JSBACH, JULES-I and ORCHIDEE-I) still include a thermally composite snow-soil layer, and the lack of a soil moisture freezing representation

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

in JSBACH causes soil temperatures to be underestimated. Although newer versions of these models (ORCHIDEE-E, ORCHIDEE-MICT, JSBACH-PF, JULES-GL7 and JULES-UKESM) include more realistic snow-soil process representations, cold biases of the implicit versions have, with the exception of ORCHIDEE-E, been replaced by warm biases, and of similar magnitude between JULES-I and JULES-GL7.

4. Discussion

4.1 Motivation behind participation

One of the motivations behind the design of ESM-SnowMIP was to run a stand-alone MIP dedicated to snow processes parallel to other MIPs, most notably CMIP6 and LS3MIP: "Combining the evaluation of these global-scale simulations with the detailed process-based assessment at the site scale provides an opportunity for substantial progress in the representation of snow, particularly in Earth system models that have not been evaluated in detail with respect to their snow parameterizations" (Krinner et al., 2018). Identifying errors in ESM-SnowMIP site simulations could be linked to model processes that also operate in LS3MIP global simulations, separately from meteorological and ancillary data errors. However, LS3MIP and ESM-SnowMIP results are not directly comparable because land surface schemes (LSSs) include parametrizations that describe sub-grid heterogeneity and some LSSs allow them to be switched off or modified for point simulations. Tables 1 and 2 show whether models participated in both MIPs and whether they used point simulation-specific snow cover parametrizations, which is critical for albedo and the most common

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

parametrization to simulate sub-grid heterogeneity. Of the eleven models that did not adjust their sub-grid parametrizations or impose complete snow cover (Figure 3), only one (CLASS) is not participating in LS3MIP. Of those that are participating, three switched off their sub-grid parametrizations (MATSIRO, RUC, and SURFEX-ISBA). Had it been anticipated at the design stage that some models would have considered ESM-SnowMIP to be a means to evaluate their LS3MIP set-up against in situ data, ESM-SnowMIP instructions would have advised to switch off all sub-grid processes; treating a point simulation like a spatial simulation makes evaluating some variables against point measurements futile. This is best illustrated with ORCHIDEE, the three versions of which have the highest negative albedo biases; not only was complete snow cover not imposed, but also the maximum albedo for deep snow on grass (i.e. 0.65 at all sites except Weissfluhjoch) accounts implicitly for sub-grid heterogeneity in large-scale simulations.

Although called ESM-SnowMIP, the site simulations were always intended to include physically based snow models that are not part of an ESM but have other applications (Krinner et al., 2018). Table 3 lists what motivated different groups to participate in ESM-SnowMIP Although not explicit in Table 3 because of the anonymity of the comments, for developers of snow physics models, the motivation to participate in a MIP dedicated to scrutinizing the processes they investigate is self-evident. On the other hand, most land surface schemes were first developed to provide the lower boundary conditions to atmospheric models. Because of the dramatic differences in the energy budget of snow-free and snow-covered land, the main requirement for snow models in some LSSs is still just to inform atmospheric models of whether there is snow on the ground or not. The size of the modelling group also matters; more models supported by a single individual or small teams listed exposure as one of their motivations. This discussion revealed that many participants suffered from the "false

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

consensus effect" (Lee et al., 1977), also observed among geoscientists but not explicitly named by Baartman et al . (2020), i.e. they assumed their motivations were universal, or at the very least, widespread. Ultimately, the prestige of MIPs means that, regardless of workload, personal motivation or model performance, they have become compulsory promotional exercises that we cannot afford not to participate in, for better or worse.

4.2 Errare humanum est

The increasing physical complexity of models makes them harder for users to understand. Many LSSs are "community" models (e.g. CLM, CoLM, JULES, SURFEX-ISBA), meaning that they are being developed and used by a broad range of scientists whose research interests, other than all being related to some aspect of the land surface, do not necessarily overlap. In many cases, new parametrizations are added faster than old ones are deprecated, causing ever-growing user interfaces or configuration files to become incomprehensible. Benchmarking should help scientists verify that newer versions of a model can reproduce the same results as older versions, but the lag between scientific improvements (hard code) and those at the user interface (soft code) can cause model errors to be introduced by simple avoidable mistakes. The JULES configuration files, for example, contain approximately 800 switches and parameters. Although GL7 and UKESM are the official JULES configurations implemented in the CMIP6 Physical Model and Earth System setups respectively, the ESM-SnowMIP results had to be re-submitted multiple times because large errors were eventually traced to a poorly documented but highly sensitive parameter.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-19-0329.1.

It should be noted that JULES and many other models were not intended for point simulations, increasing the possibility of errors in reconfiguring them for ESM-SnowMIP.

A different philosophy from some other MIPs has been followed here such that resubmission of simulations was encouraged if initial results did not appear to be representative of the intended model behaviour. Table 4 provides details of the hard- and soft-coded errors identified as a result of discussions that led to sixteen of the twenty-six models re-submitting their results, some more than once. One model was excluded at a late stage because the modelling team did not identify the source of some very large errors that caused the model to be an outlier in all analyses and, therefore, would not have added any scientific value to this paper.

Model errors can be statistically quantified; quantifying human errors is somewhat more challenging. A methodology widespread in high-risk disciplines (e.g. medicine, aviation and nuclear power), the Human Reliability Assessment, may be the closest analogue, but it is a preventative measure. Concerns about reproducibility and traceability have motivated a push for analogous methodologies in the Geosciences (Gil et al., 2016), but most remain retrospective steps to retrace at the paper writing stage.

Figure 4 quantifies the differences in the performance of the two variables (SWE and soil temperature) and models most affected by human errors before and after resubmission. For some models (JULES-GL7, JSBACH-PF, HTESSEL-ML), SWE NRMSE before resubmission are up to five times higher than after and soil temperature bias double that of corrected simulations (ORCHIDEE-I). Human errors in models and, as discussed in Menard et al. (2019) for the first ten reference sites in ESM-SnowMIP, in data are inevitable, and this snow MIP shows that they are widespread. The language we use to describe numerical models has

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

distanced them from the fact that they are not, in fact, pure descriptions of physics but rather equations and configuration files written by humans. *Errare humanum est, perseverare diabolicum.* Menard et al. (2015) showed that papers already published had used versions of JULES that included bugs affecting turbulent fluxes and causing early snowmelt. There is no requirement for authors to update papers after publication if retrospective enquiries identify some of the published results as erroneous. In view of the many errors identified here, further investigations are required to start understanding how widespread errors in publications are. Whether present in initialisation files or in the source code, these errors impair or slow progress in our understanding of snow modelling because they misrepresent the ability of models to simulate snow mass and energy balances.

4.3 Model documentation

As with many other areas of science, calls for reproducibility of model results to become a requirement for publication are gaining ground (Gil et al., 2016). Table 1 was initially intended to list the parametrizations considered most important in snow modelling (Essery et al., 2013; Essery, 2015), with, as is conventional (e.g. Rutter et al., 2009; Krinner et al., 2018), a single reference per model. Referencing the parametrizations in the twenty-seven models requires, in fact, seventy-nine papers and technical reports; a more detailed version of the table and associated references are included in the supplementary material. The lead-author first identified fifty-one references, and the modelling teams then provided references to fill the remaining gaps. However, some suggested the wrong references, others revised their initial answers and a few even discovered that some parametrizations are not described

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

at all. Not only is it extremely rare to find complete documentation of a model in a single publication, it is also difficult to find all parametrizations described at all in the literature. When this happens, some parametrizations are described in publications for other models. Often, the most recent publication refers to previous ones, which may or may not be the first to have described the model, comprehensively or not. Incomplete documentation would be an annoying but unimportant issue if this exercise had not led to the identification of some of the errors discussed in Section 4.2.

Less than a decade ago, it was at best difficult and at worst impossible to publish scientific model descriptions. The open access culture, issues of reproducibility and online platforms dedicated to publication of source code and data have reversed this trend such that it is now difficult to imagine research relying on a new model with proprietary code being published. Yet, it is a truth universally acknowledged that openly budgeting in a project proposal for the added time it takes to publish comprehensive data and model descriptions is unadvisable, despite many funding bodies enforcing open-access policies. The problem remains for models developed before the tide changed. Two examples illustrate this best. The first concerns the number of papers which refer to Anderson (1976) for snow density, liquid water retention or thermal conductivity. Equations for these parametrizations do appear in the report, but often not in the form presented in subsequent papers (Essery et al., 2012 pointed out that most actually use the forms in Jordan, 1991), or they are themselves reproductions of equations from earlier studies (especially for snow thermal conductivity). The second example is a quote taken from the paper describing VEG3D (Braun and Schädler, 2005): “The snow model is based on the Canadian Land Surface Scheme (CLASS) (Verseghy 1991) and ISBA (Douville et al. 1995) models, and accounts for changes of albedo and emissivity as well as processes like compaction, destructive metamorphosis, the melting of snow, and the freezing of liquid

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

water." This sentence is the only description in English of the snow model in VEG3D; a more comprehensive description, not referenced in Braun and Shädler (2005), is available in German in a PhD thesis (Grabe, 2002). The study in which the quote appears did not focus on snow processes, so a full description of the snow model may not have been necessary, but it is nonetheless a cause for concern that referees, at the very least, did not require clarifications as to which processes were based on CLASS and which on ISBA. Changes in emissivity certainly were not based on either model as both did – and still do – have fixed emissivity. This is the most succinct description of a snow model, but not the only one to offer little or no information about process representations. At the other end of the spectrum, the CLM5 documentation is the most comprehensive and makes all the information available in a single technical report (Lawrence et al., 2020). A few models follow closely with most information being available in a single document that clearly references where to obtain additional information (e.g. CLASS, SURFEX-ISBA, HTESSEL, JULES, SNOWPACK). The "Publish or perish" culture is estimated to foster a nine percent yearly growth rate in scientific publications (Bornmann and Mutz, 2015) which will be matched by a comparable rate of solicitations for peer reviewing. Whether it is because we do not take or have time to fact-check references, the current peer-review process is failing when poorly described models are published. The aim of LS3MIP and ESM-SnowMIP is to investigate systematic errors in models; errors can be quantified against evaluation data for any model, but poor documentation accentuates our poor understanding of model behaviour and reduces MIPs to statistical exercises rather than to insightful studies.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

5. What the future holds

Historically, PILPS (Henderson-Sellers et al., 1995) and other intercomparison projects have provided platforms to motivate model developments; they are now inextricably linked to successive IPCC reports. In view of heavily mediatised errors such as the claim that Himalayan glaciers would melt by 2035 – interestingly described as "human error" by the then IPCC chairman Rajendra Pachauri (archive.ipcc.ch, 2010; Times of India, 2010) – we must reflect on how damaging potential errors are to the climate science community. Not only are the IPCC reports the most authoritative in international climate change policy-making, but they have become – for better or worse – proxies for the credibility of climate scientists to the general public. It is therefore time that we reflect on our community and openly acknowledge that some model uncertainties cannot be quantified at present because they are due to human errors.

Other factors are also responsible for the modelling of snow processes not having progressed as fast as other areas relying on technology. Discussions on the future of snow MIPs involving organisers and participants of ESM-SnowMIP issued from this study. As in the discussion about motivation of participants, suggestions for the design of future MIPs were varied, and at times contradictory, but responses from participants reflected the purpose their models serve (Table 4). The IPCC Expert Meeting on Multi Model Evaluation Good Practice Guidance states that "there should be no minimum performance criteria for entry into the CMIP multi-model database. Researchers may select a subset of models for a particular analysis but should document the reasons why" (Knutti et al., 2010). Nevertheless, many participants argued that the "one size fits all" approach should be reconsidered. ESM-

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

SnowMIP evaluated models against the same bulk snowpack properties as previous snow MIPs. This suited LSSs that represent snow as a composite snow/soil layer or as a single layer, but there is a demand for more complex models that simulate profiles of internal snowpack properties to be evaluated against data that match the scale of the processes they represent (e.g. snow layer temperatures, liquid water content and microstructure). Models used at very high resolution for avalanche risk forecasting (such as Crocus and SNOWPACK; Morin et al., 2020) and by the tourism industry are constantly being tested during the snow season and errors can cost lives and money. However, obtaining reliable data and designing appropriate evaluation methodologies to drive progress in complex snow models is challenging (Menard et al., 2019). For example, solving the trade-off between SWE and surface temperature errors requires more measurements of surface mass and energy balance components: simple in theory but expensive and logistically difficult in practice. The scale at which even the more complex models operate is also impeding progress. Until every process can be described explicitly, the reliance of models on parametrizations to describe very small scale processes (such as the surface exchanges upon which the above trade-off depends) are inevitable sources of uncertainty.

Despite expressing a need for change in the design of snow MIPs, many participants described ESM-SnowMIP as a success because it allowed them to identify bugs or areas of their models in need of further improvements; some improvements were implemented in the course of this study, others are in development. Ultimately, ESM-SnowMIP's main flaw is of not being greater than the sum of its parts. Its working hypothesis was not supported and, per se, has failed to advance our understanding of snow processes. However, the collaborative effort allowed us to report a false, but plausible hypothesis, to expose our misplaced assumptions and to reveal a disparity of opinions on the purpose, design and future

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

of snow MIPs. In view of our findings, of the time investment required of participating modellers and of novel ways to utilise already available global-scale simulations (e.g. Mudryk et al., 2020), most planned ESM-SnowMIP experiments may not go ahead, but site simulations with evaluation data covering bulk and internal snowpack properties will be expanded. Learning from our mistakes to implement future MIPs may yet make it an unqualified success in the long term.

Acknowledgments

CM and RE were supported by NERC grant NE/P011926/1. Simulations by participating models were supported by the following programs and grants: Capability Development Fund of CSIRO Oceans and Atmosphere, Australia (CABLE); Canada Research Chairs and Global Water Futures (CRHM); H2020 APPLICATE grant 727862 (HTESSEL); Met Office Hadley Centre Climate Programme by BEIS and Defra (JULES-UKESM and GL7); TOUGOU Program from MEXT, Japan (MATSIRO); RUC by NOAA grant NOAA/NA17OAR4320101 (RUC); Russian Foundation for Basic Research grant 18-05-60216 (SPONSOR); Russian Science Foundation Grant 16-17-10039 (SWAP). ESM-SnowMIP was supported by the World Climate Research Programme's Climate and Cryosphere (CliC) core project.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

References

Abramowitz, G. and Bishop, C. H., 2015: Climate model dependence and the ensemble dependence transformation of CMIP projections, Journal of Climate, 28, 2332–2348.

Alien, 1979. [film]. Directed by R. Scott.

Anderson, E., 1976: A point energy and mass balance model of a snow cover. NOAA Tech Rep. NWS 19, 150 pp.

Archive.ipcc.ch, IPCC statement on the melting of Himalayan glaciers https://archive.ipcc.ch/pdf/presentations/himalaya-statement-20january2010.pdf, Accessed on 29 October 2019.

Baartman, J.E.M, Melsen, L.A., Moore, D. and van der Ploeg, M.J., 2020: On the complexity of model complexity: Viewpoints across the geosciences, *Catena*, 186, 10.1016/j.catena.2019.104261, https://doi.org/10.1016/j.catena.2019.104261

Bornmann, L. and Mutz, R, 2015: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references, *Journal of the Association for Information Science and Technology*, 66, 2215-2222, https://doi.org/10.1002/asi.23329

Braun, F. J. and Schädler, G., 2005: Comparison of Soil Hydraulic Parameterizations for Mesoscale Meteorological Models, J. Appl. Meteorol., 44, 1116–1132, https://doi.org/10.1175/JAM2259.1.

Clark, M. P. and Coauthors, 2015: A unified approach for process-based hydrologic modeling: 1. Modeling concept, Water Resour. Res., 51, 2498– 2514, doi:10.1002/2015WR017198.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

Conway J.P., Pomeroy J.W., Helgason W.D. and Kinar, N.J., 2018: Challenges in modelling turbulent heat fluxes to snowpacks in forest clearings, Journal of Hydrometeorology, 19, 1599–1616, https://doi.org/10.1175/JHM-D-18-0050.1.

Douville, H. and Mahfouf, J.F., 1995: A new snow parameterization for the Météo-France Climate Model. 1. Validation in stand-alone experiments. Climate Dynamics, 12, 21–35.

Essery, R., 2013: Large-scale simulations of snow albedo masking by forests, Geophys. Res. Lett., 40, 5521– 5525, doi:10.1002/grl.51008.

Essery, R., 2015: A Factorial snowpack model (FSM 1.0) Geoscientific Model Devlopemnt, 8, 3867-3876, https://doi.org/10.5194/gmd-8-3867-2015

Essery, R. and Etchevers, P., 2004: Parameter sensitivity in simulations of snowmelt, Journal of Geophysical Research, 109, 1-15, https://doi.org/10.1029/2004JD005036.

Essery, R., Morin, S., Lejeune, Y. and Menard, C., 2013: A comparison of 1701 snow models using observations from an alpine site. In: Advances in Water Resources, 55, 131-148.

Essery, R., Rutter, N., Pomeroy, J., Baxter, R., Stähli, M., Gustafsson, D., Barr, A., Bartlett, P., and Elder, K., 2009: SNOWMIP2: An Evaluation of Forest Snow Process Simulations, B. Am. Meteorol. Soc., 90, 1120–1135, https://doi.org/10.1175/2009BAMS2629.1.

Etchevers, P. and Coauthors, 2002: SnowMiP, an intercomparison of snow models: first results. In: Proceedings of the International snow science workshop, Penticton, Canada, 29 Sep.-4 Oct. 2002.

Etchevers, P. and Coauthors, 2004: Validation of the surface energy budget simulated by several snow models (SnowMIP project), Annals of Glaciology, 38, 150-158.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

Flato, G and Coauthors, 2013: Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Gil, Y. and Coauthors, 2016: Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance, Earth and Space Science, 3, 388-415, https://doi.org/10.1002/2015EA000136.

Grabe, F., 2002: Simulation der Wechselwirkung zwischen Atmosphäre, Vegetation und Erdoberfläche bei Verwendung unterschiedlicher Parametrisierungsansätze (PhD thesis, Karlsruhe University, Karlsruhe). Retrieved from KIT-Bibliothek (KITopen-ID: 122002).

Henderson-Sellers, A., Pitman, A.J., Love, P.K., Irannejad, P. and Chen, T., 1995: The project for Intercomparison of land surface parameterisaton schemes (PILPS) Phases 2 and 3. Bull. Amer. Meteor. Soc., 76, 489-503.

Jordan, R., 1991: A One-Dimensional Temperature Model for a Snow Cover, Technical Documentation for SNTHERM.89, Special Report 91-16, U.S. Army Corps of Engineers, 62 pp.

Kerr, N. L., 1998: HARKing: hypothesizing after the results are known. Personality and Social Psychology Review, 2, 196–217.

Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P.J., Hewitson, B. and Mearns L., 2010: Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections. In: Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections [Stocker, T.F., D. Qin,

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

G.-K. Plattner, M. Tignor, and P.M. Midgley (eds.)]. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland.

Knutti, R., Sedláček, J, Sanderson, B.M., Lorenz, R., Fischer, E.M. and Eyring, V, 2017: A climate model projection weighting scheme accounting for performance and interdependence. Geophysical Research Letters, 44, 1909–1918.

Krinner, G. and Coauthors, 2018: ESM-SnowMIP: Assessing models and quantifying snow-related climate feedbacks, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-153.

Lafaysse, M., Cluzet, B., Dumont, M., Lejeune, Y., Vionnet, V., and Morin, S., 2017: A multiphysical ensemble system of numerical snow modelling, The Cryosphere, 11, 1173-1198, https://doi.org/10.5194/tc-11-1173-2017.

Lawrence, D. and Coauthors, 2020: Technical Description of version 5.0 of the Community Land Model (CLM), NCAR, 329pp, http://www.cesm.ucar.edu/models/cesm2/land/CLM50_Tech_Note.pdf

Lee, R., Greene, D. and House, P., 1977: The 'false consensus effect': An egocentric bias in social perception and attribution processes, Journal of Experimental Social Psychology, 13, 279–301. https://doi.org/10.1016/0022-1031(77)90049-X.

Li, Y., Wang, T., Zeng, Z., Peng, S., Lian, X and Piao S., 2016: Evaluating biases in simulated land surface albedo from CMIP5 global climate models, Journal of Geophysical Research: Atmospheres, 121, 6178-6190, https://doi.org/10.1002/2016JD024774

Martin, G.R.R., 1996: A Game of Thrones, Bantam Spectra, USA.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

Menard, C.B., Ikonen, J., Rautiainen, K., Aurela, M., Arslan, A.N., and Pulliainen, J., 2015: Effects of Meteorological and Ancillary Data, Temporal Averaging, and Evaluation Methods on Model Performance and Uncertainty in a Land Surface Model. J. Hydrometeor., 16, 2559–2576, https://doi.org/10.1175/JHM-D-15-0013.1.

Menard, C. B. and Coauthors, 2019: Meteorological and evaluation datasets for snow modelling at 10 reference sites: description of in situ and bias-corrected reanalysis data, Earth Syst. Sci. Data, 11, 865–880, https://doi.org/10.5194/essd-11-865-2019.

Morin S. and Coauthors, 2020: Application of physical snowpack models in support of operational avalanche hazard forecasting: A status report on current implementations and prospects for the future, Cold Regions Science and Technology, 170, 102910, https://doi.org/10.1016/j.coldregions.2019.102910

Mudryk, L., Santolaria-Otín, M., Krinner, G., Ménégoz, M., Derksen, C., Brutel-Vuilmet, C., Brady, M., and Essery, R., 2020: Historical Northern Hemisphere snow cover trends and projected changes in the CMIP-6 multi-model ensemble, The Cryosphere Discuss., https://doi.org/10.5194/tc-2019-320.

Munafò, M. and Coauthors, 2017: A manifesto for reproducible science, Nature Human Behaviour, 1, 0021, https://doi.org/10.1038/s41562-016-0021.

Niu, G.-Y. and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. J. Geophys. Res., 116, D12109, doi: 10.1029/2010JD015139.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

Pitman, A.J. and Henderson-Sellers, A., 1998: Recent progress and results from the project for the intercomparison of land surface parameterization schemes. Journal of Hydrology, 212-213, 128-135, https://doi.org/10.1016/S0022-1694(98)00206-6

Psycho, 1960: [film] Directed by A. Hitchcock.

Randall, D.A. and Coauthors, 2007: Climate Models and Their Evaluation. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M. and Miller, H.L. (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Roesch, A., 2006: Evaluation of surface albedo and snow cover in AR4 coupled climate models, J. Geophys. Res., 111, D15111, doi:10.1029/2005JD006473.

Rutter, N. and Coauthors, 2009: Evaluation of forest snow processes models (SnowMIP2), J. Geophys. Res., 114, D06111, https://doi.org/10.1029/2008JD011063.

Sandu, I., Beljaars, A., Bechtold, P., Mauritsen, T., and Balsamo, G., 2013: Why is it so difficult to represent stably stratified conditions in numerical weather prediction (NWP) models?, J. Adv. Model. Earth Syst., 5, 117– 133, doi:10.1002/jame.20013.

Slater, A.G. and Coauthors, 2001: The representation of snow in land surface schemes: results from PILPS 2(d). Journal of Hydrometeorology, 2, 7–25.

https://timesofindia.indiatimes.com/videos/news/Pachauri-admits-mistake-in-IPCC-report/videoshow/5492814.cms, 2010, Pachauri admits mistake in IPCC report, accessed on 4 October 2019.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

van den Hurk, B. and Coauthors, 2016: LS3MIP (v1.0) contribution to CMIP6: the Land Surface, Snow and Soil moisture Model Intercomparison Project – aims, setup and expected outcome, Geosci. Model Dev., 9, 2809-2832, https://doi.org/10.5194/gmd-9-2809-2016.

Verseghy, D. L., 1991: Class—A Canadian land surface scheme for GCMS. I. Soil model. Int. J. Climatol., 11, 111-133, https://doi.org/10.1002/joc.3370110202.

Wang, L., Cole, J.N.S., Bartlett, P., Verseghy, D., Derksen, C., Brown, R., and von Salzen, K., 2016: Investigating the spread in surface albedo for snow-covered forests in CMIP5 models, Journal of Geophysical Research: Atmospheres, 121, 1104– 1119, doi:10.1002/2015JD023824.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

*Table 1: Key characteristics of snow model parametrizations and variables on which they depend, and number of papers per model over which descriptions of the seven parametrizations are spread. Abbreviations and symbols: LWC = Liquid water content, SCF = snow cover fraction ("point" means models used point-specific parametrizations, "grid" means they did not), MC = Mechanical compaction, OL = Obukhov length, PC = Personal communication, Ri_b = bulk Richardson number, * = references provided by personal communication and cannot be traced in the existing literature about this specific model. A more detailed version of this table including full references for parametrizations is available in the supplementary material.*

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

<table>
<tr><th></th><th>Albedo</th><th>Conductivity</th><th>Density</th><th>Turbulent fluxes</th><th>LWC</th><th>SCF</th><th>Snow layering</th><th>n Papers</th></tr>
<tr><td>CABLE-SLI</td><td>Spectral</td><td>Power function</td><td>MC</td><td>OL</td><td>Yes</td><td>Point</td><td>Single</td><td>3</td></tr>
<tr><td>CLASS</td><td>Spectral</td><td>Quadratic equation</td><td>Time</td><td>Ri_B</td><td>Yes</td><td>Grid</td><td>Single</td><td>2</td></tr>
<tr><td>CLM5</td><td>Spectral</td><td>Density</td><td>MC</td><td>OL</td><td>Yes</td><td>Grid</td><td>Multi</td><td>1</td></tr>
<tr><td>CoLM</td><td>Spectral</td><td>Quadratic equation</td><td>MC</td><td>OL</td><td>Yes</td><td>Grid</td><td>Multi</td><td>7*</td></tr>
<tr><td>CRHM</td><td>Spectral</td><td>Density and humidity</td><td>MC</td><td>OL</td><td>Yes</td><td>Point</td><td>Multi</td><td>4* + PC</td></tr>
<tr><td>Crocus</td><td>Spectral</td><td>Power function</td><td>MC</td><td>Ri_B.</td><td>Yes</td><td>Point</td><td>Multi</td><td>3</td></tr>
<tr><td>EC-EARTH</td><td>Time and temperature</td><td>Power function</td><td>MC</td><td>OL</td><td>Yes</td><td>Grid</td><td>Single</td><td>3*</td></tr>
<tr><td>ESCIMO</td><td>Temperature</td><td>None</td><td>Time</td><td>Empirical</td><td>Yes</td><td>Point</td><td>Single</td><td>3*</td></tr>
<tr><td>HTESSEL</td><td rowspan="2">Time and temperature</td><td rowspan="2">Power function</td><td rowspan="2">MC</td><td rowspan="2">OL</td><td rowspan="2">Yes</td><td rowspan="2">Grid</td><td>Single</td><td>3</td></tr>
<tr><td>HTESSEL (ML)</td><td>Multi</td><td>3</td></tr>
<tr><td>SURFEX-ISBA</td><td>Spectral</td><td>Power function</td><td>MC</td><td>Ri_B.</td><td>Yes</td><td>Point</td><td>Multi</td><td>2</td></tr>
<tr><td>JSBACH</td><td rowspan="2">Spectral</td><td>Fixed</td><td>Fixed</td><td rowspan="2">OL</td><td rowspan="2">No</td><td rowspan="2">Point</td><td>Composite</td><td>3*</td></tr>
<tr><td>JSBACH3-PF</td><td>Power function</td><td>Time</td><td>Multi</td><td>4*</td></tr>
<tr><td>JULES-GL7 JULES-UKESM</td><td>Spectral</td><td>Power function</td><td>MC</td><td>OL</td><td>Yes</td><td>Point</td><td>Multi</td><td>2</td></tr>
<tr><td>JULES-I</td><td>Temperature</td><td>Fixed</td><td>Fixed</td><td>OL</td><td>No.</td><td>Point</td><td>Composite</td><td>1</td></tr>
<tr><td>MATSIRO</td><td>Spectral</td><td>Fixed</td><td>Fixed</td><td>OL</td><td>No.</td><td>Point</td><td>Multi</td><td>3</td></tr>
<tr><td>ORCHIDEE-E ORCHIDEE-MICT</td><td rowspan="2">Time</td><td>Quadratic equation</td><td>MC</td><td rowspan="2">OL</td><td>Yes</td><td>Grid</td><td>Multi</td><td>1 + PC</td></tr>
<tr><td>ORCHIDEE-I</td><td>Fixed</td><td>Fixed</td><td>No</td><td>Grid</td><td>Composite</td><td>3 + PC</td></tr>
<tr><td>RUC</td><td>Time</td><td>Fixed</td><td>MC</td><td>OL</td><td>No</td><td>Grid</td><td>Multi</td><td>3 + PC</td></tr>
<tr><td>SMAP</td><td>Spectral</td><td>Quadratic equation</td><td>MC</td><td>OL</td><td>Yes</td><td>Point</td><td>Multi</td><td>3</td></tr>
<tr><td>SNOWPACK</td><td>Statistical</td><td>Conductivity model</td><td>Empirical</td><td>OL</td><td>Yes</td><td>Point</td><td>Multi</td><td>5</td></tr>
<tr><td>SPONSOR</td><td>Time</td><td>Density</td><td>MC</td><td>OL</td><td>Yes</td><td>Grid</td><td>Multi</td><td>2 + PC</td></tr>
</table>

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

SWAP	Density	Density	SWE and snow	OL	Yes	Point	Single	3
VEG3D	Time	Density	Time	OL	No	Point	Single	4*

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

Table 2: Participating models and modelling teams. ESM-SnowMIP provided vegetation height, soil type and snow-free albedo to the participants; where relevant, these may differ from LS3MIP configurations.

Model	ESM-SnowMIP contact	Model type	Model version	Model configuration	Differences between LS3MIP and ESM-SnowMIP configurations
CABLE-SLI	Matthias Cuntz, Vanessa Haverd	LSS in Access	CABLE revision 4252	CABLE including SLI as described in Haverd and Cuntz (2016). Snow and ice extensions as in Cuntz and Haverd (2018). 12 soil layers.	Did not participate in LS3MIP
CLASS	Paul Bartlett	LSS in CanESM	CLASS 3.6.2	CLASS-CTEM off-line code with CTEM turned off, and using the 2-band snow albedo and associated snow-ageing scheme. Initialization files are available on demand. Other than adjustments to match the site properties (e.g. soil type, vegetation, snow-free albedo) all parameters are the model default values.	Did not participate in LS3MIP
CLM5	Sean Swenson	LSS in CESM	CLM5.0	Standard	No difference.
CoLM	Yongjiu Dai, Hua Yuan	LSS in BNU-ESM and CAS-ESM	CoLM Version 2014	Default	CoLM Version 2005 Many differences including pedotransfer functions of soil hydraulic and thermal parameters, numerical solution of Richards

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

					equation of soil water content.
CRHM	Xing Fang, John Pomeroy	Hydrological model	CRHM 01/17/18	Adapted from CRHM plot-scale simulation project for coniferous forest and forest clearing sites in Canadian Rocky Mountains detailed in Pomeroy et al. (2012) with modified configuration for soil module allowing simulations for permafrost and seasonal frost.	Did not participate in LS3MIP
Crocus	Matthieu Lafaysse	Snow physics model	Git tag ESM-SnowMIP-Crocus-ESCROC (= commit b57f02d6 4/12/2017)	Crocus : default configuration as defined in Lafaysse et al. (2017), Figure 2. Drift module allowing change of physical properties of near surface snow activated for SNB and WFJ.	Did not participate in LS3MIP
EC-EARTH	Emanuel Dutra	LSS in EC-EARTH	EC-EARTH v3.2.2 revision r4381	Offline "OSM" configuration with prescribed surface albedo and vegetation.	LS3MIP simulation will be done with the latest "frozen" model version for CMIP6, including interactive vegetation and variable surface albedo.
ESCIMO	Thomas Marke, Ulrich Strasser	Snow surface energy balance model	ESCIMO v5 based on ESCIMO v4 with additional functionality described in Marke et al. (2016).	Albedo parameterization as in Cox et al. (1999) Sensible heat equation as in Weber (2008) Empirical density function as in Essery et al. (2013)	Did not participate in LS3MIP
HTESSEL HTESSEL-ML	Gabriele Arduini	LSS of ECMWF operational forecasting system	HTESSEL cycle 43r3	Operational HTESSEL configuration uses the single layer snow scheme from Dutra et al. (2010). The experimental HTESSEL configuration (HTESSELML)	Did not participate in LS3MIP

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

				uses a multi-layer snow scheme documented in Arduini et al. 2019 (under review in JAMES). Note that the configuration of the multi-layer snow scheme and model cycle used for ESM-SnowMIP runs differ from Arduini et al. (2019).	
SURFEX-ISBA	Bertrand Decharme, Aaron Boone	LSS in CNRM-CM	SURFEX version 8.0 (ISBA and all related schemes including snow are embedded in the SURFEX numerical platform)	As in Decharme et al. (2016) denoted as the "NEW" experiment.	Snow grid-cell fraction doesn't account for vegetation in the 1-dimensional ESM-SnowMIP runs.
JSBACH3 JSBACH3-PF	Stefan Hagemann	LSS in MPI-ESM	JSBACH3 (Revision 9168, state of 31.07.2017) and JSBACH3-PF (same revision but with improved snow parametrizations inherited from JSBACH4)	Time step: 450s, With YASSO soil model, no dynamic vegetation, no nitrogen, no disturbances and no land use transitions. Orography and LAI do not affect surface roughness. Soil states were initialized from previous global offline simulation using GWSP3 forcing. JSBACH3-PF uses the "permafrost" configuration with enabled soil freezing and thawing, and with related processes based on Ekici et al. (2014).	JSBACH-PF did not participate in LS3MIP JSBACH3: No difference
JULES-I	Cecile Menard, Richard Essery	LSS in HadCM3	JULES 4.8 (Revision 7629)	Zero-layer snow model as described in Best et al. (2011).	Did not participate in LS3MIP
JULES-GL7 JULES-UKESM	Eleanor Burke	LSS in HadGEM3-GC3 and UKESM	JULES 5.3	GL7 and UKESM configurations with site-specific characteristics.	Different fractional snow cover parametrization for plot-scale and distributed simulations.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

MATSIRO	Tomoko Nitta, Hyungjun Kim	LSS in MIROC	MATSIRO 6	MATSIRO for offline land simulations. The configuration is the same as the GSWP3 simulations except for subgrid-scale parameterizations (tile scheme, SSNOWD snow cover parameterization and arctic wetland scheme), which are turned off for plot-scale simulations.	All subgrid-scale parameterizations are tuned off for plot-scale simulations.
ORCHIDEE-E ORCHIDEE-I ORCHIDEE-MICT	Claire Brutel-Vuilmet, Gerhard Krinner	LSS in IPSL-CM	ORCHIDEE E and I TRUNK revision 4695; ORCHIDEE MICT 8.7.1 revision 5308	TRUNK is the version of ORCHIDEE that is used in the first CMIP6 runs. We have the implicit snow version (TRUNK-I) which is the older snow that was used in CMIP5 and the explicit snow version (TRUNK-E) that is used in CMIP6 (based on Wang et al., 2013). MICT is the high-latitude version of ORCHIDEE (Guimberteau et al., 2018).	No difference.
RUC	Tatiana Smirnova	LSS in NOAA/NCEP operational forecasting systems	RUC model – WRF 4.0 official release	Standard RUC configuration for offline simulations: 9 levels in soil, 2-layer snow model with separate treatment of snow-covered and snow-free areas for patchy snow.	Subgrid-scale parameterizations for fractional snow cover and surface parameters are turned off for ESM-SnowMIP.
SMAP	Masashi Niwano	Snow physics model	SMAP v4.23rc1	SMAP v4.23rc1	Did not participate in LS3MIP
SNOWPACK	Nander Wever, Charles Fierz	Snow physics model	MeteoIO preprocessing library: revision 2011 from https://models.slf.ch/svn/meteoio/trunk SNOWPACK model:	The standard version of SNOWPACK was used, in default configuration.	Did not participate in LS3MIP

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

			revision 1480 from https://models.slf.ch/svn/snowpack/branches/dev		
SPONSOR	Dmitry Turkov, Vladimir Semenov	Hydrological model	SPONSOR, ver.2.0	The model was adapted for calculations of spatially distributed landscape characteristics with observed meteorological forcing. The latest version of the snow model is described in Turkov and Sokratov (2016).	No difference
SWAP	Olga Nasonova, Yeugeny Gusev	LSS	As described in Gusev and Nasonova (2003)	As described in Gusev and Nasonova (2003)	Did not participate in LS3MIP
VEG3D	Gerd Schädler	Soil and vegetation model	As described in Braun and Schädler (2005)	Standard configuration: 8 soil layers, time step 300 s.	Did not participate in LS3MIP

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

Motivation behind participation	Future of snow MIPS
• To identify key missing processes. • To cut out the noise from ensemble simulations in order to extract the signal. • To compare how models implement snow processes and, if possible, what are the implications. • To have a detailed analysis of one's own model; doing the model simulations is easier than analysing the results. • To provide new insights into modelling. • To document the current state of the models. • To help modellers understand their and other models better. • To determine the skill of an operational model in offline simulations before starting coupled simulations for weather predictions. • To motivate model improvements. • To participate in the beauty contest (the statistical performance of my not-so-sophisticated model is similar to complex process-based models). • To identify a range of "good enough" - models reflecting the range of process uncertainty.	• Allow re-submission of simulations if errors are identified. • Provide model code and initialisation files as well as model results for transparency. • Move towards a more process-based diagnostic in order to improve parametrizations and not just to tune parameters. • Need new evaluation metrics. • Evaluate against internal snowpack properties (e.g. snow layer thermal conductivity, temperature, density). • Move towards fewer models with multiple hypotheses (e.g. FSM, Essery, 2015; SUMMA, Clark et al., 2015; or Noah-MP, Niu et al., 2011) • Cluster models depending on their complexity. • Not all models should be accepted. There could be minimum requirements in terms of parametrizations (e.g. stability dependent exchange coefficients); outliers from the previous experiment would not be allowed to participate in the next stages; new models should present a proof of energy and moisture conservation in their models.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

• To make one's model visible to the snow modelling community. • To be part of the snow modelling community. • To evaluate one's model at reference sites across different elevation gradients and climatic settings. • To avoid equifinality problems by evaluating models performance with multiple variables that contribute to and are relevant to snow processes. • To provide benchmarks against which to evaluate models.	• All models should be accepted, but different levels of involvement should be allowed so modelling groups can choose the experiments they want to participate in. • Constrain model sensitivity with observations (e.g. SWE, snow albedo) or fixed variables. • Provide evaluation data at the same time as the forcing data. • Provide fewer sites as initialisation of many sites can be a source of human errors. • Provide more challenging sites (e.g. tundra, wind-blown).

Table 3: Summary of discussions with ESM-SnowMIP participants about (1) what motivated them to participate and (2) their suggestions about the design of the next snow MIP.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

Table 4: Hard and soft coded errors identified by the results analysis team (AT) or modelling team (MT) in the course of this study.

	Unusual model behaviour	**Model**
Soft-coded errors		
Did not change start time between SNB and SWA (start at 00:00) and other sites (start at 01:00)	Mismatched timestamps (AT)	All models
Initial conditions taken from wrong date	Mismatched timestamps (AT)	CLM5
Specified site-specific parameters not taken from site descriptions	Unrealistically low albedo with consequences on snow mass and melt (AT)	JSBACH, JSBACH-PF, JULES-I
Wrong forcing file used for one site	Models results were identical at two sites (AT)	RUC
Simulations used UTC times instead of local times	Unrealistically high albedo (AT)	Crocus
Many variations in output file formats; wrong variable name; variations in the interpretation of the ESM-SnowMIP definition of output variables; different sign conventions.	N/A	One or more in adjacent list for most, if not all, models.
Errors in converting to ESM-SnowMIP format because of the above.	N/A	Some models. Results analysis team.
Hard-coded errors		
Bug in model use of site longitude	Unrealistically low albedo with consequences on snow mass and melt (AT)	JULES-GL7, JULES-UKESM

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

Bug in transmission of SW radiation through canopy	Investigated slow melting behaviour of model after evaluation data became available (MT)	SURFEX-ISBA
Model SWE limited to a maximum of 1000 mm	SWE limited to 1000 mm (AT)	MATSIRO
Unintentional decoupling of snow surface and atmosphere	Snow did not melt at Weissfluhjoch in some summers (AT)	HTESSEL-ML
Bug in partitioning of SW radiation into direct and diffuse	Unrealistically high albedo values (AT)	Crocus
Bug in the output of liquid water content	Found unrealistically small liquid water content values when compared ESM-SnowMIP results with other simulations (MT)	HTESSEL, EC-EARTH
Inconsistent use of snow area fraction when calculating snow depth and SWE	Snow density varied instead of being fixed (AT)	MATSIRO
Many variations in output file formats; wrong variable name; variations in the interpretation of the ESM-SnowMIP definition of output variables; different sign conventions.	N/A	One or more in adjacent list for most, if not all, models.
Errors in converting to ESM-SnowMIP format because of the above.	N/A	Some models. Results analysis team.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

Figures

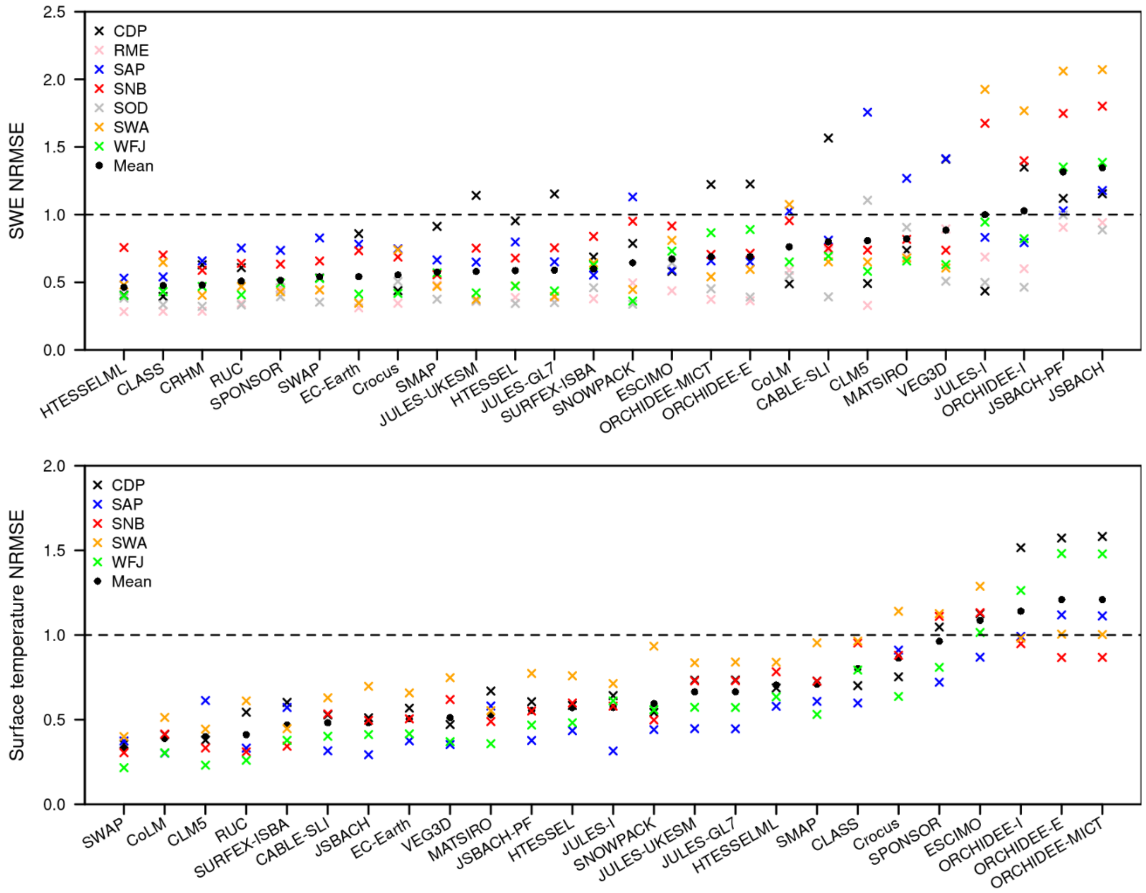

Figure 1: Model ranking by normalised root mean square errors of snow water equivalent and surface temperature. The site names are shortened as follows: CDP = Col de Porte, SAP = Sapporo, RME = Reynolds Mountain East, SNB = Senator Beck, SOD = Sodankylä, SWA = Swamp Angel and WFJ = Weissfluhjoch.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-19-0329.1.

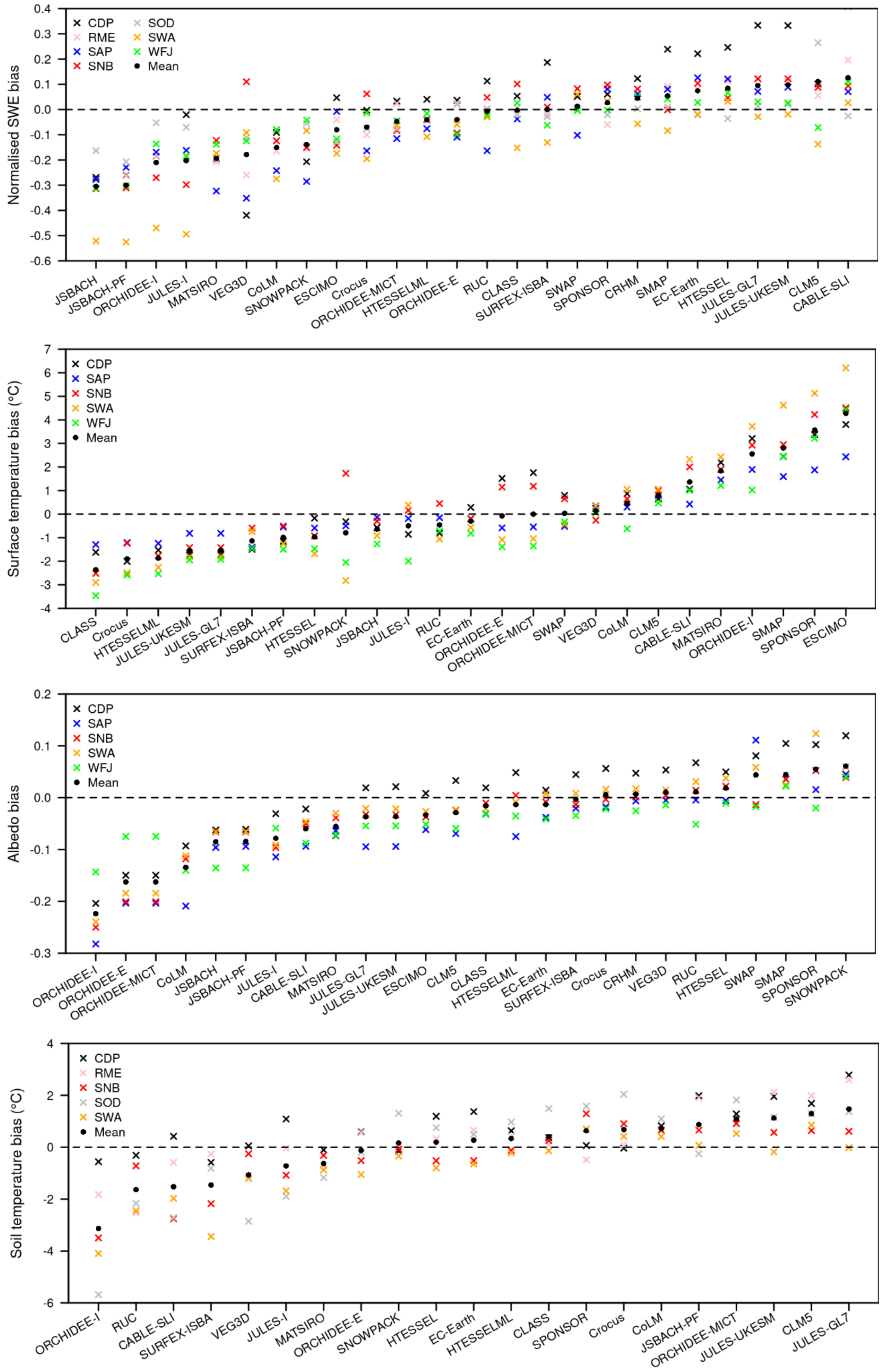

Figure 2: Model ranking by biases from negative to positive. Following the prevalent

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

convention, negative biases denote model underestimates. SWE biases are normalised by measured mean yearly maxima. JSBACH soil temperature cold biases (ranging from -6°C to -12°C and averaging -9°C) are outside the range of the plot. The site names are shortened as in Figure 1.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

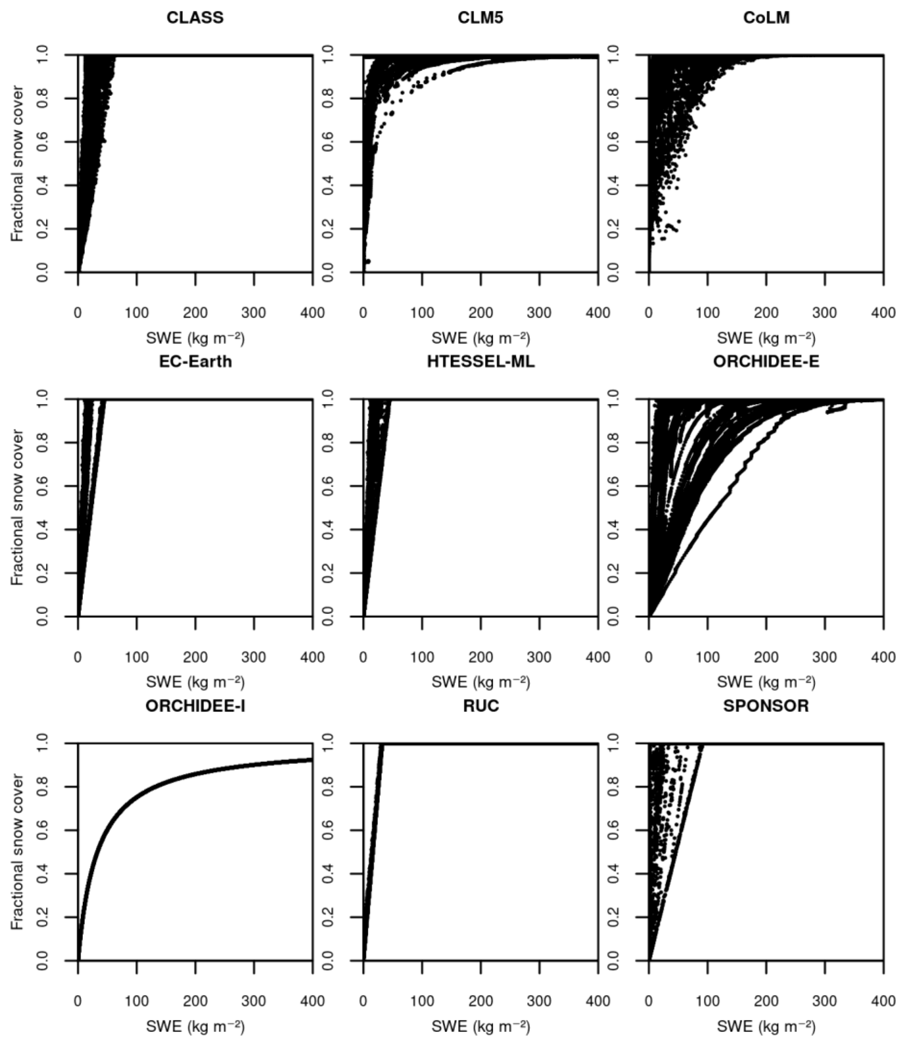

Figure 3: Fractional snow cover (SCF) as a function of SWE at Col de Porte for models that did not switch off their sub-grid parametrizations or impose complete snow cover. HTESSEL is not shown as it is the same as HTESSEL-ML. ORCHIDEE-MICT did not force SCF = 1, but values were missing from the file provided for evaluation.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.

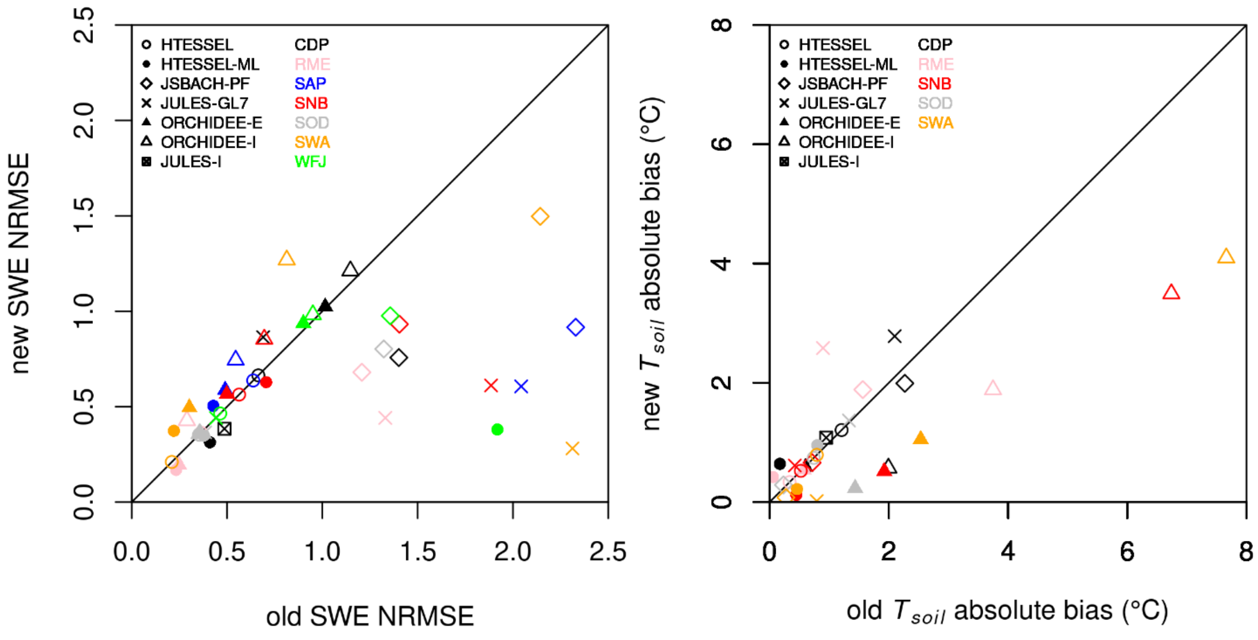

Figure 4: SWE NRMSE and soil temperature (T_{soil}) absolute bias before and after resubmission for selected models. The site names are shortened as in Figure 1.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-19-0329.1.