ELEMENTA
Science of the Anthropocene

Solomon, A, et al. 2023. The winter central Arctic surface energy budget: A model evaluation using observations from the MOSAiC campaign. *Elem Sci Anth*, 11: 1. DOI: https://doi.org/10.1525/elementa.2022.00104

RESEARCH ARTICLE

The winter central Arctic surface energy budget: A model evaluation using observations from the MOSAiC campaign

Amy Solomon[1,*], Matthew D. Shupe[1], Gunilla Svensson[2], Neil P. Barton[3,4], Yurii Batrak[5], Eric Bazile[6], Jonathan J. Day[7], James D. Doyle[3], Helmut P. Frank[8], Sarah Keeley[7], Teresa Remes[5], and Mikhail Tolstykh[9]

This study evaluates the simulation of wintertime (15 October, 2019, to 15 March, 2020) statistics of the central Arctic near-surface atmosphere and surface energy budget observed during the MOSAiC campaign with short-term forecasts from 7 state-of-the-art operational and experimental forecast systems. Five of these systems are fully coupled ocean-sea ice-atmosphere models. Forecast systems need to simultaneously simulate the impact of radiative effects, turbulence, and precipitation processes on the surface energy budget and near-surface atmospheric conditions in order to produce useful forecasts of the Arctic system. This study focuses on processes unique to the Arctic, such as, the representation of liquid-bearing clouds at cold temperatures and the representation of a persistent stable boundary layer. It is found that contemporary models still struggle to maintain liquid water in clouds at cold temperatures. Given the simple balance between net longwave radiation, sensible heat flux, and conductive ground flux in the wintertime Arctic surface energy balance, a bias in one of these components manifests as a compensating bias in other terms. This study highlights the different manifestations of model bias and the potential implications on other terms. Three general types of challenges are found within the models evaluated: representing the radiative impact of clouds, representing the interaction of atmospheric heat fluxes with sub-surface fluxes (i.e., snow and ice properties), and representing the relationship between stability and turbulent heat fluxes.

Keywords: MOSAiC, Arctic wintertime boundary layer statistics, Coupled forecast systems, Surface energy budget, Model evaluation, Model intercomparison

1. Introduction

To produce useful forecasts and projections of the Arctic system, it is necessary to accurately simulate the transfer of energy through the coupled ocean-sea ice-snow-atmosphere system. This transfer of energy happens at the snow/sea ice/ocean–atmosphere interface and is described physically by the surface energy balance (SEB). The SEB is a function of the surface storage term, surface radiation, turbulent flux, and conductive ground flux. To simulate this balance, and the evolution of the Arctic system, it is necessary to adequately simulate the processes that determine the evolution of the atmospheric boundary layer, such as, cloud processes, turbulence, conduction through snow and sea ice, and the coupling between these processes.

It is challenging to evaluate and improve the simulation of Arctic boundary layer processes, especially in the central Arctic in winter due, in part, to limited observations. A notable exception is the Surface Heat Budget of the Arctic campaign (SHEBA; Moritz et al., 1993; Perovich et al., 1999; Uttal et al., 2002), a year-long drift experiment that took place in the Beaufort and Chukchi Seas from October 1997 to October 1998. Here we provide a brief overview of how SHEBA observations have been used to identify biases in simulations of the Arctic system as

[1] Cooperative Institute for Research in Environmental Sciences, University of Colorado and NOAA Physical Sciences Laboratory, Boulder, CO, USA
[2] Department of Meteorology, Stockholm University, Stockholm, Sweden
[3] Naval Research Laboratory, Monterey, CA, USA
[4] Current address: NOAA Center for Weather and Climate Prediction, College Park, MD, USA
[5] Development Centre for Weather Forecasting, Norwegian Meteorological Institute, Oslo, Norway
[6] CNRM – UMR3589, Météo-France and CNRS, Toulouse, France
[7] European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom
[8] Deutscher Wetterdienst, Research and Development, Offenbach/Main, Germany
[9] Marchuk Institute of Numerical Mathematics RAS and Hydrometcentre of Russia, Moscow, Russia

* Corresponding author:
Email: amy.solomon@noaa.gov

background for this current study. Observations taken during SHEBA have been used to identify wintertime biases in regional climate models (Tjernstroüm et al., 2005; Rinke et al., 2006; Wyser et al., 2008), and global forecast systems (Beesley et al., 2000; Brunke et al., 2006; Simjanovski et al., 2011).

A significant finding from these studies was that 5 out of 6 regional climate models were unable to simulate the decrease in turbulence for very stably stratified boundary layers due to a constant diagnosed transfer coefficient in the sensible heat flux parameterization (Tjernstroüm et al., 2005). In addition, in very stable conditions the radiative flux divergence near the surface may be as important as the sensible heat flux divergence (see Kondo et al., 1978) and may require increased model resolution near the surface. It was found that errors in the surface heat flux were often at least as large as the net heat flux itself. Five of the models produced 2–4 m s^{-1} near-surface wind speed errors during wintertime (Tjernstroüm et al., 2005). In addition, in an evaluation of wintertime flux algorithms, Brunke et al. (2006) found that, while fluxes for the unstable regime were generally within the range of observed values, sensible heat flux for stably stratified boundary layers was an order of magnitude too large in 3 of the models, with only one algorithm adequately simulating the decrease in turbulent heat flux for bulk Richardson numbers greater than 0.25.

The recent studies of Sedlar et al. (2020) and Inoue et al. (2020) demonstrate that current state-of-the-art regional climate model simulations over the Arctic Ocean constrained with fixed surface conditions and atmospheric nudging still have large biases in atmospheric boundary layer and cloud processes, specifically the phase partitioning between cloud liquid water and ice, and the resultant impact on the atmospheric boundary layer (e.g., Køltzow et al., 2019; Tjernstroüm et al., 2021). Clearly, climate models and forecast systems are still challenged to simulate a number of fundamental aspects of the Arctic climate system.

A recent year-long drift experiment called the Multidisciplinary drifting Observatory for the Study of Arctic Climate expedition (MOSAiC; see Shupe et al., 2022), which took place from October 2019 to October 2020, was designed to address these challenges. The stated goal of the MOSAiC campaign is to enhance understanding of central Arctic coupled atmosphere-ice-ocean-ecosystem processes to improve numerical models for sea ice forecasting, extended-range weather forecasting, climate projections, and climate change assessment. MOSAiC took place 22 years after SHEBA over thinner and more dynamic sea ice. Similar to SHEBA, MOSAiC was a drift experiment, where an icebreaker was docked next to an ice floe, thereafter the icebreaker and constellation of instruments surrounding the ship drifted with the ice floe over an annual cycle. Both campaigns took intensive measurements of all components of the surface energy budget, as well as, atmospheric boundary layer, cloud, snow, sea ice, and ocean properties (among other components) at a number of different sites. However, MOSAiC had sites 10–20 km from the icebreaker in order to make observations at multiple locations across a domain the size of a "floating climate model grid box." Also, different from SHEBA, the MOSAiC campaign took place on the Atlantic side of the Arctic, where strong winds accelerate the Trans-polar drift of the sea ice and its transport through Fram Strait, and warm, moist air masses from the North Atlantic are advected into the Arctic. SHEBA took place in the Pacific sector of the Arctic Ocean, with less influence from lower latitudes and slower ice drift. Also, during MOSAiC, the sea ice approximately 40 km around the Central Observatory was residual first-year ice (Krumpen et al., 2021), while SHEBA took place over multiyear sea ice. However, observations of sea ice thickness during SHEBA ranged from 0.3 to 8 m and a wide variety of ice types was observed (Perovich et al., 2003). Also, both SHEBA and MOSAiC took place during anomalous atmospheric conditions; SHEBA took place during the 1997–1998 ENSO event and MOSAiC experienced the strongest polar vortex on record after February 2020 (Lawrence et al., 2020).

In this study we take advantage of these new observations to assess model simulations of the Arctic system from a coupled perspective in order to evaluate the atmospheric processes that impact the surface energy budget and feedback to the ocean and sea ice. It is only in a coupled system that models can simultaneously simulate the impact of cloud-driven radiative effects, turbulence, and precipitation processes on the surface layer and growth of the sea ice. Wintertime observations (here defined as the period without solar radiation during MOSAiC, 15 October, 2019—15 March, 2020) are used to evaluate coupled processes unique to the Arctic, such as: The representation of liquid-bearing clouds at cold temperatures; the representation of a persistent stable boundary layer; and the limiting impact of atmospheric variability on sea ice by snow. Short-term forecasts are used in this study to identify potential errors in the representation of "fast" processes, such as cloud feedbacks and surface fluxes, that cause biases in climate model projections of Arctic climate change. The relative importance of these processes in the models is studied from the perspective of the surface energy budget.

The MOSAiC campaign and observations used in this study are described in Section 2. Section 3 describes the models used in this study. Results of the model intercomparison are presented in Section 4. In Section 5 we discuss findings and present our conclusions.

2. Observations used in this study

MOSAiC observations started in October 2019 when the Alfred Wegener Institute icebreaker, R/V *Polarstern* (Knust, 2017), was docked along an ice floe in the North Laptev Sea. Krumpen et al. (2020) showed that the sea ice within 40 km of the ship was generally younger and thinner than surrounding ice and it was formed in a polynya event north of the New Siberian Islands at the beginning of December 2018. They determined that those sea ice conditions were due to the interplay between a high ice export in the late winter preceding MOSAiC and high air temperatures during the following summer, which yielded

Table 1. MOSAiC observations used in this study

Instrument	Measurements	Location	Reference
GNDRAD: radiometers	Upwelling broadband longwave and shortwave surface radiation	Met City	Sengupta et al. (2021) doi:10.5439/1025192
ICERAD: radiometers	Downwelling broadband longwave and shortwave surface radiation	Met City	Riihimaki (2021) doi:10.5439/1608608
10-m tower: sonic anemometer	Temperature, winds, relative humidity, sensible heat flux at nominally 10-m	Met City	Cox et al. (2021d) doi:10.18739/A2VM42Z5F
ASFS: radiometers, sonic anemometer	Temperature, relative humidity at nominally 2-m. Winds and sensible heat flux at 3.8 m. Upwelling and downwelling broadband longwave and shortwave surface radiation	L-sites	Cox et al. (2021a, 2021b, 2021c) doi:10.18739/A20C4SM1J doi:10.18739/A2CJ87M7G doi:10.18739/A2445HD46
Combined: Ka-band ARM Zenith Radar, Microwave radiometers, Micropulse lidar, ceilometer, radiosonde	Cloud liquid and ice water paths derived from multiple measurements and a cloud classification algorithm	Polarstern	Shupe (2022) doi:10.5439/1871015

the longest ice-free summer. For Siberian shelf seas this resulted in the longest ice-free summer since the beginning of observations, 93 days.

Polar night (here defined as 15 October–15 March) started shortly after the instruments were set up. A storm passed over the MOSAiC site after the campaign was completely set up in mid-November, bringing high winds that caused leads to form throughout the camp and observational network. Twelve cyclones were observed during the Polar night. These cyclones advected warm, moist air masses from lower latitudes toward the MOSAiC location (Rinke et al., 2021). For example, in late February, a storm passed over the MOSAiC site producing persistent strong winds and a mid-winter warming to −10°C. Prolonged quiescent periods were also observed, with high pressure systems during late-December and early-March producing weak winds and extreme cold temperatures, down to −42.3°C on March 4, 2020. More context for the atmospheric variability is discussed in the MOSAiC atmosphere overview paper (Shupe et al., 2022).

A Central Observatory (CO) was set up on the MOSAiC floe within 2 km of the *Polarstern*. The CO included a "Met City" where a 10-m and 30-m tower were set up to measure temperature, relative humidity, and winds at different heights, along with a number of other measurement systems. A distributed network was set up with numerous semi-autonomous stations and buoys up to 25 km from the *Polarstern*. Three sites with comprehensive measurements of the ocean/ice/atmosphere system were called "L-sites" and were initially positioned 13–23 km from *Polarstern* at nominally 120° intervals. Using measurements from the L-sites and the Met City together allows for estimates of variability on the scale of a climate model grid box. Using the range of observed values in the comparison against models addresses the question of how representative the CO is of the area covered by the gridbox of the models. The constellation of the *Polarstern*, the CO, and the distributed network drifted freely from mid-October until mid-May when the *Polarstern* had to leave the MOSAiC floe to exchange crew and resupply in Svalbard. Therefore, MOSAiC was passively drifting for the entire period of this study, 15 October, 2019 to 15 March, 2020. This period includes all of Legs 1 and 2 and the first 3 weeks of Leg 3.

MOSAiC observed all components of the climate system (ecosystem, biogeochemistry, ocean, atmosphere, cryosphere). Of interest to this study are coincident measurements of the snow-ABL-cloud physical systems that were measured at the Central Observatory/Met City and the 3 L-sites. Surface characteristics and surface energy budget observations were taken by Atmospheric Surface Flux Stations (ASFS) that measured all components of the surface energy budget at the 3 L sites (Cox et al., 2021a, 2021b, 2021c; Shupe et al., 2022) and a 10-m tower (Cox et al., 2021d) and broadband radiation suite (Riihimaki, 2021) at Met City. Cloud properties, including the liquid water path (LWP) and ice water path (IWP), were derived from a multisensor approach that combined Ka-band Cloud Radar, microwave radiometer, micropulse lidar, ceilometer, and radiosonde observations (Shupe et al., 2015; Shupe, 2022) all made onboard *Polarstern*. See **Table 1** for details about the MOSAiC observations used in this study.

Winds at the 3 L-sites were measured at a height of 3.8 m. The Monin–Obukhov similarity theory (Monin and Obukhov, 1954) is used to scale winds to 10 m for comparison to models. Following the boundary layer scheme used in the CICE model (Hunke and Lipscomb, 2008), stability functions from Kauffman and Large (2002) are used for unstable boundary layers and stability functions from Jordan et al. (1999) are used for stable boundary layers. Using measurements from the Tower as a test, the scaling increases winds linearly as speeds increase, for example, 13 ms^{-1} wind speeds at 4 m are increased to 14.5 ms^{-1} at 10 m. Sensible heat flux at 10 m is assumed to be the same as at 3.8 m based on the assumption of constant flux in the surface layer. Hourly averaged values

Table 2. Forecast systems used in this study

Forecast System (Abbreviation Used) (Operational or Experimental)	Max Lead Times	Initial Hour	Output Frequency (Averaging)
CAFS (experimental)	2 days	0Z	Hourly (averaged except LWP/IWP)
IFS (operational)	2 days	0Z	12 min (averaged over model time step)
ARPEGE-GELATO (ARPEGE) (operational)	10 days	0Z	Hourly (instant, fluxes averaged)
SL-AV (operational)	7 days	0Z	15 min (instant, fluxes averaged)
ICON (DWD) (operational)	7.5 days	0Z	Hourly (average over 2 min model time step)
HARMONIE-AROME version cy43h1.2 (H-AROME) (experimental)	2.5 days	0Z	Hourly (instant, fluxes averaged)
Navy-ESPC (NAVY) (operational)	2 days	12Z	1–3 hourly (instantaneous)

Table 3. Domains, resolution, and sea ice/snow models used in the different forecast systems

Forecast System	Domain	Atmosphere Horizontal Resolution (km)	Lowest Atmospheric Model Level (m)	Sea Ice/Snow Model
CAFS	Pan-Arctic	~10	12	Sea ice and snow model with 7 ice layers and 1 snow layer (CICE4). Hunke and Lipscomb (2008)
IFS	Global	~18	9	1.5-m sea ice thickness. No snow on sea ice. Keeley and Mogensen (2018)
ARPEGE	Global	7.5	9	Sea ice and snow model with 10 ice layers and 1 snow layer (GELATO). Salas Mélia (2002)
SL-AV	Global	~18	29	Specified sea ice. No snow on sea ice.
DWD	Global	~13	10	Sea ice model with 1 ice layer. Snow represented with empirical temperature dependence. Mironov et al. (2012)
H-AROME	Central Arctic	2.5	11	1D sea ice model with 4 ice layers and 12 snow layers. Batrak et al. (2018); Batrak and Muller (2019)
NAVY	Global	37	10	Sea ice and snow model with 7 ice layers and 1 snow layer (CICE4). Hunke and Lipscomb (2008)

are used in this study. The MOSAiC dataset has a cadence of 10 min and hourly averages at each site are set to missing if any 10 min average within the hour is missing.

3. Models used in this study

Seven different operational and experimental forecast systems are used in this model intercomparison study; the NOAA-PSL Coupled Arctic Forecast System (CAFS; Solomon et al., 2023), the ECMWF Integrated Forecast System (IFS; Haiden et al., 2019), the Météo-France ARPEGE-GELATO forecast system (ARPEGE; Bazile et al., 2020), the Russian Hydrometcentre SL-AV forecast system (SL-AV; Tolstykh et al., 2018), the German Weather Service forecast system (DWD; Zängl et al., 2015), the experimental configuration of the HARMONIE-AROME forecast system (H-AROME; Bengtsson et al., 2017), and the U.S. Navy-ESPC forecast system (NAVY; Barton et al., 2021). Details for the forecast systems are provided in **Tables 2** and **3**. Modeling centers provided timeseries for this study at the grid point closest to the location of the MOSAiC Central Observatory or the 4 grid points surrounding the *Polarstern*, which are used to interpolate to the *Polarstern* location.

Since the majority of these model systems are used for operational forecasting, each modeling center contributed forecasts with different maximum lead times, cadences of the model output, averaging time, and for one model, the hour of initialization. Fortunately, all model output except for the Navy-ESPC model, which has 3-hourly output after the first 12 h, can be sampled hourly. Each model has different output frequencies and averaging procedures; however, it is possible to use hourly averaged fluxes for all models except one although the number of samples in averages varies. In addition, it is possible to use instantaneous variables for a number of atmospheric fields for all models except one. These differences need to be taken into consideration when interpreting the results. Also, the

consistent lead time across all model output provided for this study is 2 days. Therefore, diagnostics in this study use hourly output for the first 2 days in order to have the same number of samples from each model and focus primarily on biases in statistics for the winter season.

Statistics used in this study are primarily based on probability distribution functions (PDFs) and cumulative distribution functions (CDFs). The statistics of each distribution are described in terms of mode, mean, median, skew, and kurtosis calculations. A mode is defined as a local maximum. CDFs are used to calculate the cumulative frequency of occurrence for a range of values, for example, the cumulative frequency of occurrence when downward longwave radiation is less than a threshold value. The mean of the distribution is calculated as $\mu = \sum xf(x)$, where x is the variable and $f(x)$ is the probability distribution. The median is calculated as the value of the field where the CDF is 50%. The standard deviation of the distribution is calculated as $\sigma = \sqrt{\sum (x-\mu)^2 f(x)}$. Skew (S) is a measure of the distributional asymmetry, typically referred to as the weight of the tails, and is the third standardized moment of the distribution. It is calculated as $S = \sum (x-\mu)^3 f(x)/\sigma^3$. A skewed distribution is not Gaussian. A negatively skewed PDF has mode > median > mean (heavier left tails), while a positively skewed distribution has mean > median > mode (heavier right tails). Kurtosis (K) is a measure of the sharpness of the distribution and is calculated as $K = \sum (x-\mu)^4 f(x)/\sigma^4$. For a Gaussian distribution K = 3. The excess kurtosis (KE) is then KE = K − 3. KE less than zero indicates a distribution that has a flatter distribution than a normal distribution, while KE greater than zero indicates a distribution that is sharper than a normal distribution with less values in the tails relative to the peak. Estimates for all of these moments are sensitive to the sample size and number of bins used in the PDFs. Since this study uses a relatively small sample size (153 forecasts, each with 48 hourly samples), the statistics used in this study are estimates that provide insight into the different representation of processes in the models.

4. Results

4.1. Surface energy budget in the wintertime Arctic

An objective of this study is to assess how model biases in turbulent and radiative fluxes impact the net surface energy budget. To do this it is necessary to develop diagnostics that show the biases in these processes and the relationships between processes, since this determines the evolution of the surface energy. Assuming a negligibly thin snow/ice layer at the interface with the atmosphere (storage is negligible for averages greater than an hour), the steady state surface energy budget (SEB) is

$$\mathrm{SH} + \mathrm{LH} - \mathrm{RADNET} = \mathrm{COND}, \tag{1}$$

where COND is the net conductive surface flux (or hereafter conductive flux), RADNET is the net surface radiative flux, SH is sensible heat flux, and LH is latent heat flux (using the standard definitions, SH, LH, and COND positive upward and RADNET positive downward). This balance assumes there is no freezing or melting of ice/snow at the interface. In the Arctic in wintertime when there is no solar radiation and latent heat flux is small, this equation can be approximated with

$$\mathrm{COND} \approx \mathrm{SH} - \mathrm{LWNET}. \tag{2}$$

LWNET is equal to the downward surface longwave flux (LWD) minus the upward surface longwave flux (LWU). Where LWU is equal to:

$$\mathrm{LWU} = \varepsilon\sigma T_s^4, \tag{3}$$

where ε is the bulk emissivity, σ is the Stefan-Boltzmann constant, and T_s is the surface skin temperature.

As will be demonstrated throughout this article, LWNET is one of the more important controlling factors for the wintertime Arctic surface energy budget over sea ice. We therefore start the evaluation of the forecast systems with LWNET, then the components of LWNET; LWD together with liquid and ice water paths (LWP and IWP), since cloud properties control the surface downward longwave radiation; LWU together with surface temperature, since the 2 are directly related. We then focus on the variables that are used by sensible heat flux parameterizations (surface wind speed and near-surface stratification) before evaluating sensible heat flux biases. After that we focus on 3 process-oriented diagnostics to evaluate the biases in the components of the SEB from a process perspective, and the relationship between the components of the SEB.

4.2. Surface longwave radiation biases

Probability Density Functions (PDFs) of wintertime LWNET for the MOSAiC campaign and the 7 forecast models at the *Polarstern* location are shown in **Figure 1**. The observed range across the 4 MOSAiC sites is shown with gray shading in all PDFs. The observed distributions show the well-known bimodal distribution in the Arctic with a peak in the −30 to −50 Wm^{-2} range for the clear-sky regime, where outgoing longwave radiation exceeds downward longwave radiation, and a second peak in the range of −10 to 0 Wm^{-2} for the cloudy sky regime, where there is near equilibration between cloud-emitted downward longwave radiation and surface-emitted upward longwave radiation (see Persson et al., 2002; Stramler et al., 2011). For this study, the "clear-sky" mode of this distribution is defined as net longwave radiation less than −25 Wm^{-2}, which includes some very thin clouds that have very little impact on atmospheric radiation. The opaque cloudy sky state is defined as LWNET greater than −15 Wm^{-2}. This state is typically composed of optically thick clouds that contain liquid water (Shupe and Intrieri, 2004), but can also include deep cloud ice layers with heavy snowfall. The thin cloud regime is defined as LWNET between −25 and −15 Wm^{-2}, and typically includes clouds with little to no liquid water, but enough condensed mass to weakly impact atmospheric radiation reaching the surface. Using Gaussian functions for the clear-sky and cloudy modes separately, the thin cloud regime is where the cloudy distribution goes to zero to

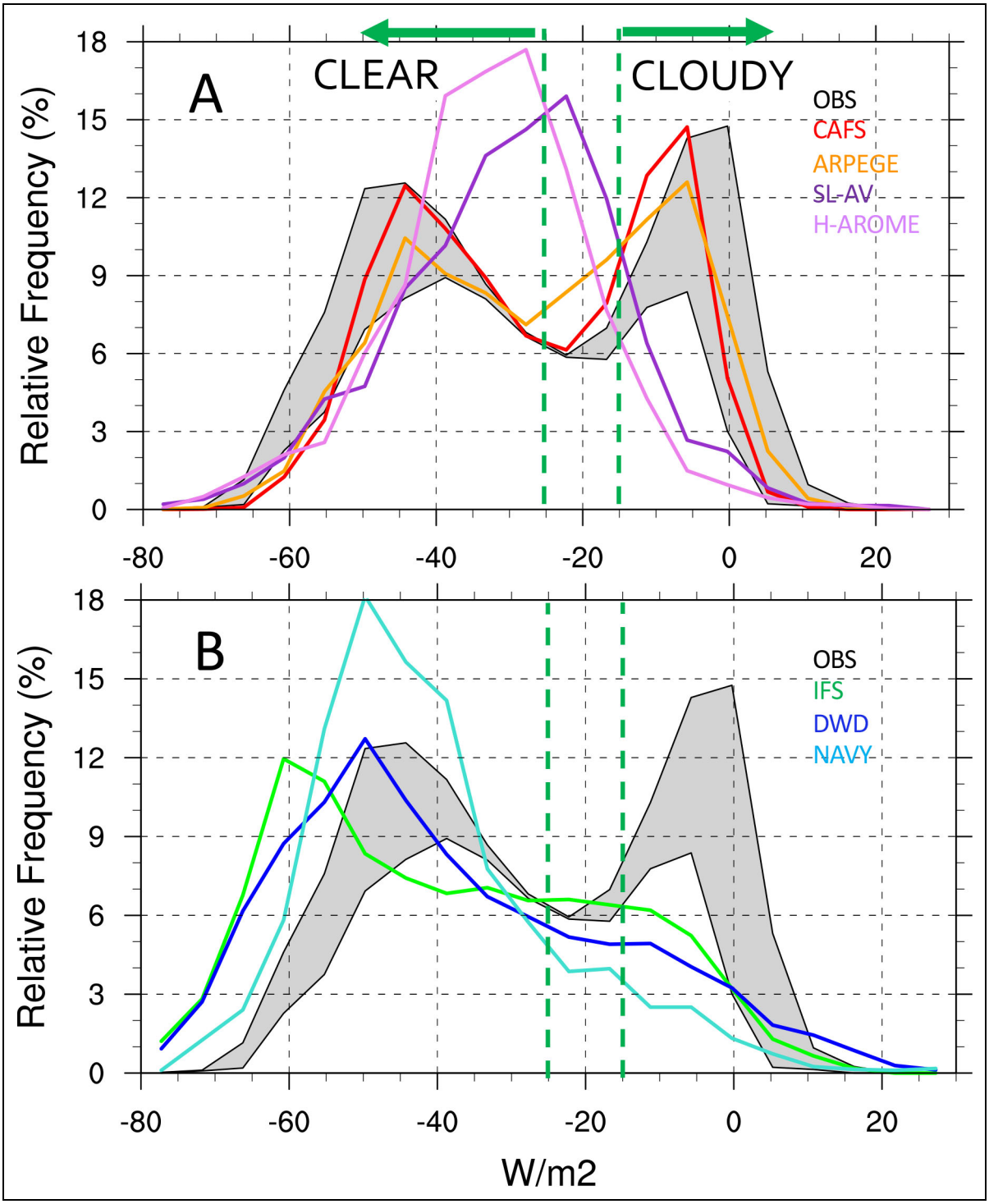

Figure 1. Probability Distribution Functions (PDFs) of model hourly averaged net surface longwave fluxes, in units of Wm^{-2}, using 1 h to 2 day lead times. The gray shading in both plots shows the range of the observed distributions using hourly averaged measurements from the 4 MOSAiC sites. Individual model distributions (shown with colored lines as indicated in the legends) are separated into 2 plots in order to clearly distinguish differences; (A) CAFS, ARPEGE, H-AROME, SL-AV; (B) IFS, DWD, NAVY. Clear-sky periods are defined in this article as net longwave fluxes less than −25 Wm^{-2}. Cloudy sky periods defined as net longwave fluxes greater than −15 Wm^{-2}.

approximately 2 standard deviations and explains 30% of the cloudy occurrences (results not shown). These cloud state definitions are used for the observations and the models in the following analysis.

PDFs of LWNET from the 7 forecast systems using hourly averaged output for 0–2 day lead times from daily forecasts relative to observations are shown with colored lines in **Figure 1**. For clarity, all figures showing PDFs are

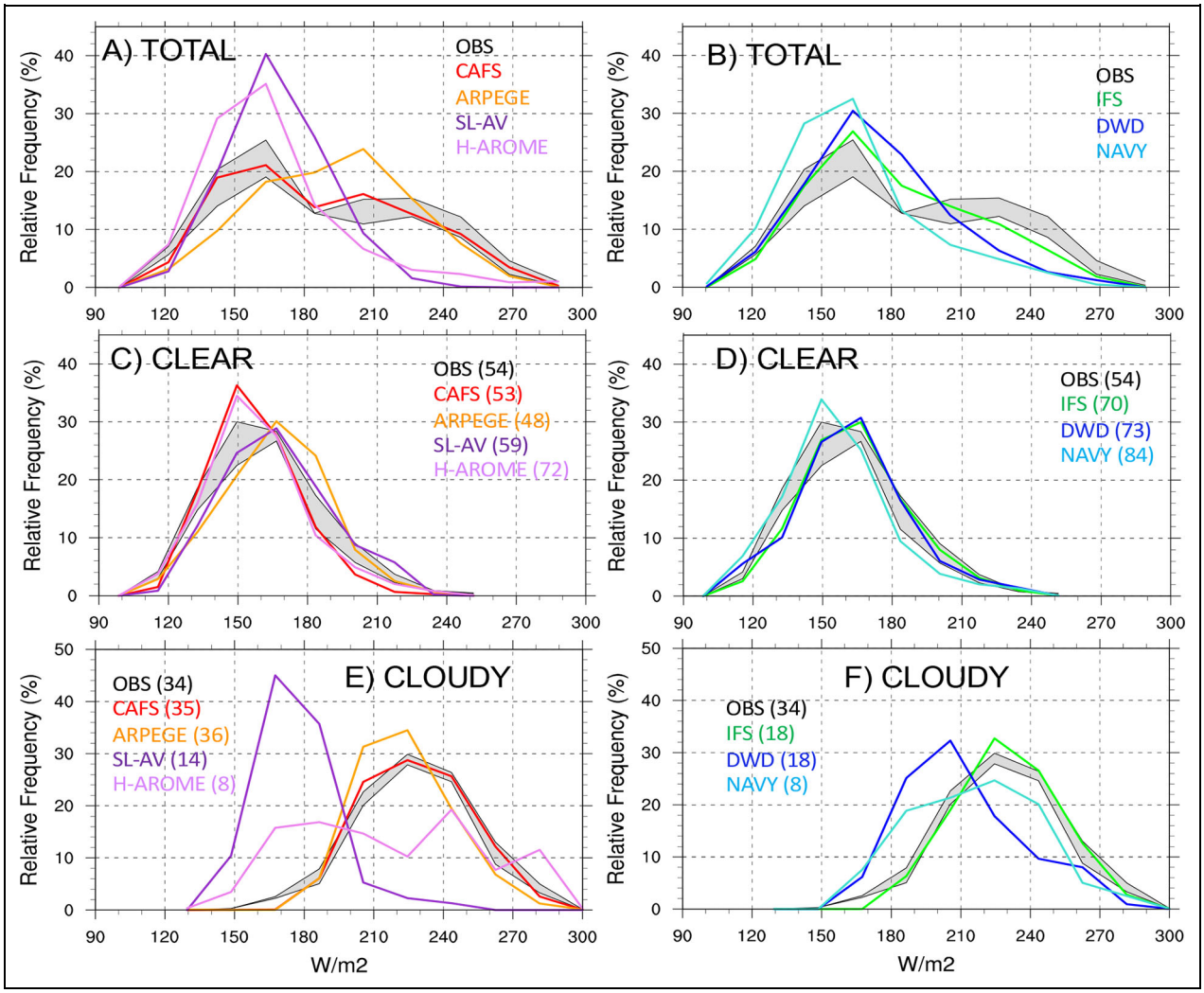

Figure 2. Probability Distribution Functions (PDFs) of model hourly averaged downward surface longwave fluxes, in units of Wm^{-2}, using 1 h to 2 day lead times. (A and B) Using all hourly samples. (C and D) Using hourly clear-sky samples. (E and F) Using hourly cloudy samples. The gray shading shows the range of the observed distributions using hourly averaged measurements from the 4 MOSAiC sites. Individual model distributions shown with colored lines as indicated in the legends. The numbers in the legends for the clear and cloudy distributions show the percent of the total sample used in the PDFs.

divided into 2 panels. It is difficult to see the variability with all model PDFs on one figure. Therefore, all top figures show the 2 models with the bimodal LWNET PDF and the 2 models with the unimodal LWNET PDF between the 2 observed modes. All bottom figures show the 3 models that underestimate the cloudy mode but have a clear-sky mode close to observations. **Figure 1A** shows that only 2 models have a bimodal distribution similar to observations (CAFS and ARPEGE) and 2 models have a unimodal distribution with a peak between the observed clear-sky and cloudy modes (SL-AV and H-AROME). **Figure 1B** shows that 3 of the models have a more dominant clear-sky mode and an underestimate of the cloudy mode (IFS, DWD, and NAVY). In addition to underestimating the cloudy distribution, these 3 models produce clear-sky distributions shifted toward larger negative LWNET magnitudes, while the distributions shown in **Figure 1A** have less negative skew than MOSAiC observations, indicating fewer clear-sky values less than -50 W m^{-2} but these distributions are within the observed range.

Looking at downward longwave flux (LWD) and upward longwave flux (LWU) separately (**Figures 2** and **3**) reveals some of the processes that cause the differences in LWNET across the models. The numbers in the legends for the clear and cloudy distributions show the percent of the total sample used in the PDFs. The observed distribution of LWD shows a bi-modal distribution with a clear-sky mode that peaks at 165 Wm^{-2} and a cloudy mode that peaks at 225 Wm^{-2} (**Figure 2A** and **B**). The observed distribution of LWU is unimodal with a peak at -200 Wm^{-2} and is negatively skewed (**Figure 3A** and **B**). This skew is primarily due to the clear-sky occurrences with the coldest surface temperature (**Figure 3A** and **C**). Interestingly, even though CAFS and ARPEGE have similar LWNET distributions, the LWD and LWU distributions show

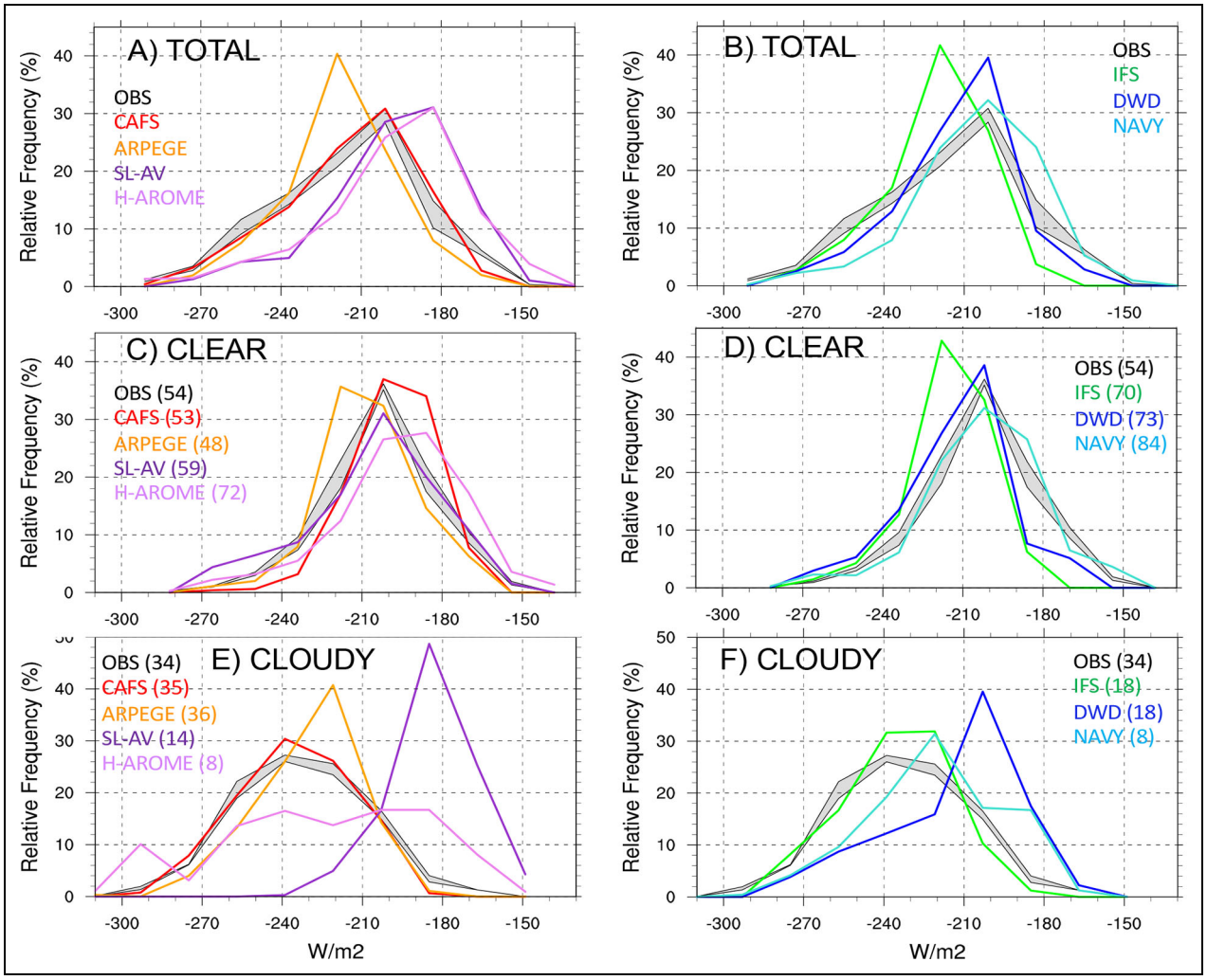

Figure 3. Probability Distribution Functions (PDFs) of model hourly averaged upwelling surface longwave fluxes, in units of Wm^{-2}, using 1 h to 2 day lead times. (A and B) Using all hourly samples. (C and D) Using hourly clear-sky samples. (E and F) Using hourly cloudy samples. The gray shading shows the range of the observed distributions using hourly averaged measurements from the 4 MOSAiC sites. Individual model distributions shown with colored lines as indicated in the legends. The numbers in the legends for the clear and cloudy distributions show the percent of the total sample used in the PDFs.

notable differences. A ~20 Wm^{-2} overestimate in the magnitude of both ARPEGE LWD and LWU clear-sky events results in LWNET values close to CAFS and the observations. In **Figure 2A**, CAFS and ARPEGE have a cloudy mode with LWD that is shifted to values 15–20 Wm^{-2} lower than observed. However, looking at **Figure 2C** and **E**, it is clear that the shift in the ARPEGE LWNET distribution is due to a clear-sky distribution that is shifted to larger values and a cloudy distribution that is shifted toward smaller values, which produces a unimodal distribution between the 2 observed modes.

4.3. Relation between LWNET biases and water paths

The relation between LWNET biases and water paths can be seen in **Figure 4**, which shows the number of occurrences for liquid and ice water paths greater than 10 gm^{-2} in 7 gm^{-2} bins, LWP and IWP, respectively. Observed maximum and minimum 1 min averages within each hour are shown with grey shading in both panels. **Figure 4A** shows that all models except one dramatically underestimate the maintenance of liquid in clouds at cold temperatures. CAFS, IFS, and SL-AV produce realistic LWP occurrences in the 13 gm^{-2} bin (**Figure 4A**). CAFS and IFS produce realistic LWP occurrences in the 20 gm^{-2} bin, however, all models that underestimate the cloudy mode underestimate LWP in bins greater than 20 gm^{-2}. Only CAFS has realistic LWP for bins greater than 20 gm^{-2}. **Figure 4B** shows that all models produce too much cloud ice during the MOSAiC winter season. However, all models except NAVY produce too many ice clouds with less than 20 gm^{-2}. CAFS and IFS produce realistic distributions up to approximately 150 gm^{-2}. ARPEGE produces too much cloud ice in all bins. This explains how LWNET can be similar to observations and CAFS even though this model does not maintain liquid in clouds at cold temperatures.

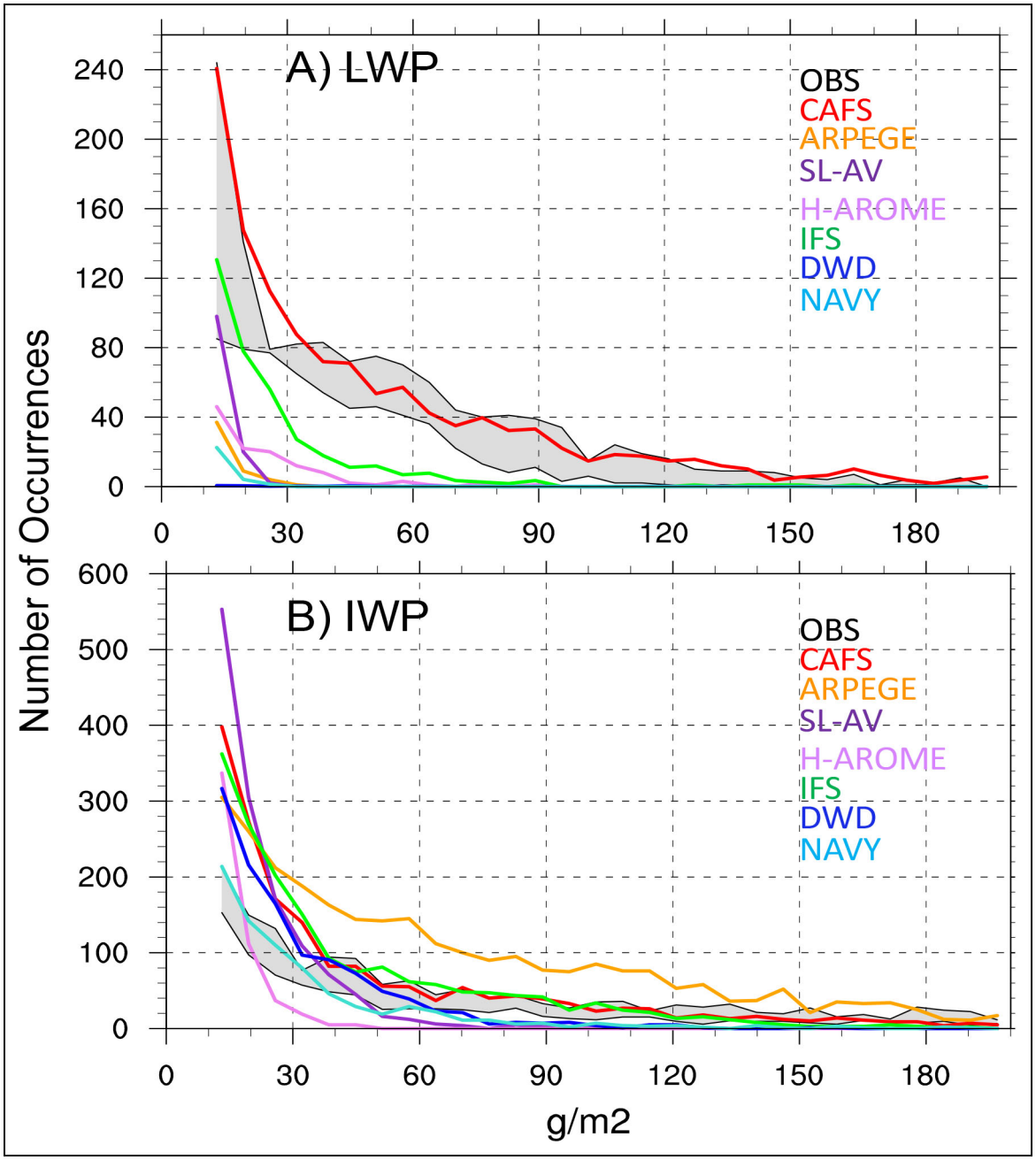

Figure 4. Number of occurrences of modeled (A) LWP greater than 10 gm^{-2} and (B) IWP greater than 10 gm^{-2} over the winter season. The gray shading shows the observed range within each hour using 1 min averages. Individual model distributions using the first 24 hours of each simulation shown with colored lines as indicated in the legends.

4.4. Relation between surface radiation biases and surface temperature

In this section, we evaluate surface temperature biases in relation to the longwave radiation results shown in **Figures 1–3**. Consistent with the LWU PDFs, CAFS has a slight shift toward colder temperatures, which shifts the median temperature by −1°C, primarily due to biases in clear-sky occurrences. ARPEGE median surface temperature is shifted high by 2°C, again primarily due to clear-sky occurrences (**Figure 5**). The CAFS LWU distribution closely approximates the observed distribution but looking at clear and cloudy occurrences separately shows that CAFS clear-sky surface emission tends to be too low (surface too cold) relative to the observations (**Figure 5**). The ARPEGE LWU distribution is sharply peaked (has less negative skew and positive KE relative to observations), in other words, has fewer extreme events and 10% more cases with LWU less than −210 Wm^{-2} relative to observations. The results together indicate that ARPEGE produces an overestimate of thin clouds, an underestimate of clear-skies, and a surface temperature distribution that essentially removes the observed positive skew toward warm temperatures and shifts the modal peak by 5.5°C (**Figure 5A**). Both CAFS and ARPEGE do a reasonable job

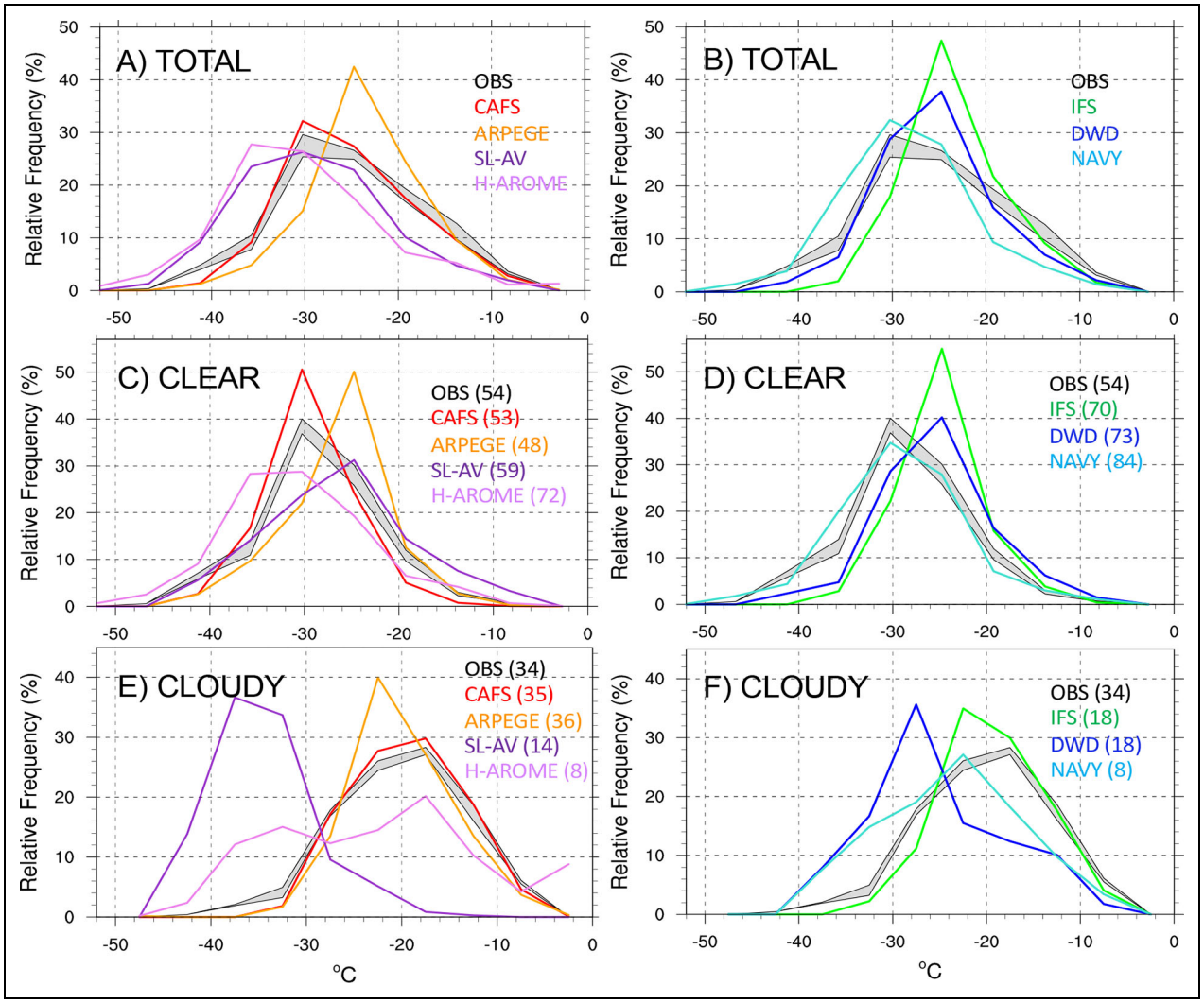

Figure 5. Probability Distribution Functions (PDFs) of model hourly averaged surface temperature, in units of °C, using 1 h to 2 day lead times. (A and B) Using all hourly samples. (C and D) Using hourly clear-sky samples. (E and F) Using hourly cloudy samples. The gray shading shows the range of the observed distributions using hourly averaged measurements from the 4 MOSAiC sites. Individual model distributions shown with colored lines as indicated in the legends. The numbers in the legends for the clear and cloudy distributions show the percent of the total sample used in the PDFs. Note difference in the observed distribution from **Figure 2** is due to a smaller number of bins used in this figure.

simulating the skew and kurtosis of the surface temperatures for cloudy conditions, with a shift in median temperatures of −1°C for ARPEGE and −0.2°C for CAFS.

The SL-AV and H-AROME biases in LWNET can be explained as the absence of a cloudy mode. Specifically, LWD greater than 200 Wm^{-2} includes less than 10% of the cases while in the observations the cumulative frequency of occurrence is closer to 35% and a ∼20 Wm^{-2} decrease in magnitude of the LWU mode peak (**Figures 2E** and **3E**), indicating that surface temperatures are too cold (**Figure 5A**). This shift is primarily due to a shift in the clear-sky occurrences in H-AROME and a shift in the cloudy occurrences for SL-AV. SL-AV and H-AROME simulate the skew and kurtosis of the total surface temperature distribution but the distributions are shifted by −4°C and −6°C, respectively. The bias in H-AROME is due to shifts toward colder temperatures in both clear-sky and cloudy occurrences.

All the models shown in **Figure 1B** show a shift in the clear-sky mode toward larger magnitudes relative to observations, but the shift in the NAVY distribution is due to 10% more cases with LWD less than 150 Wm^{-2} than the observations, while the shift in the DWD and IFS LWNET clear-sky distributions are due to 10% fewer cases with LWD less than 160 Wm^{-2} together with 20% more cases with LWU less than −210 Wm^{-2} than the observations. In other words, the NAVY model bias is primarily due to clear-sky LWD being too small, while DWD and IFS have clear-sky LWD that is too large, but this is compensated in the LWNET by surface temperatures that are too warm (**Figure 5D**). All 3 models in **Figure 5B** underestimate the frequency of the cloudy mode but have different surface temperature distributions, with the NAVY median temperature being 4°C too low, IFS being 2°C too high, and DWD being very close to observations. IFS does a good

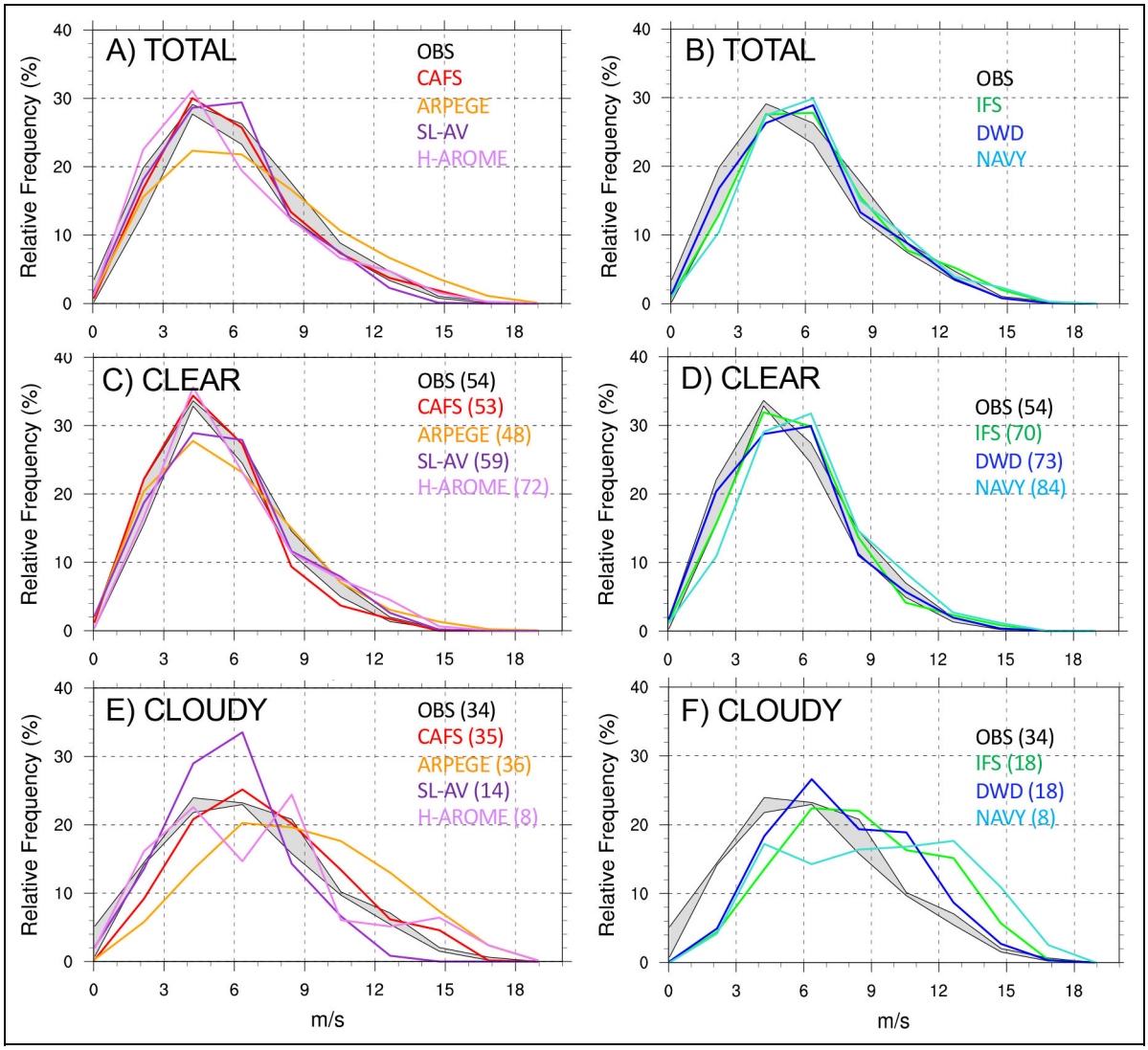

Figure 6. Probability Distribution Functions (PDFs) of model hourly averaged 10-m wind speed, in units of ms^{-1}, using 1 h to 2 day lead times. (A and B) Using all hourly samples. (C and D) Using hourly clear-sky samples. (E and F) Using hourly cloudy samples. The gray shading shows the range of the observed distributions using hourly averaged measurements from the 4 MOSAiC sites. Individual model distributions shown with colored lines as indicated in the legends. The numbers in the legends for the clear and cloudy distributions show the percent of the total sample used in the PDFs.

job of simulating the skew and kurtosis of the observed cloudy distribution, while DWD and NAVY have cloudy distributions that overestimate the frequency of occurrence for surface temperatures lower than −24°C by 32% and 22%, respectively.

4.5. Wind speed biases

Figure 6 shows the tower 10-m wind speed PDFs for all-sky, clear-sky, and cloudy distributions. Looking at the all-sky distributions first (**Figure 6A** and **B**), all models except ARPEGE closely approximate the observed 10-m wind speed distribution. ARPEGE has heavier tails than observations, meaning more occurrences of extreme high wind speeds, resulting in a median wind speed that is 1 ms^{-1} larger than observations. All models produce realistic distributions of clear-sky wind speed. Observed 10-m distributions show that the median value for cloudy skies (5.5 ms^{-1}) is higher than the value found for clear-sky occurrences (4 ms^{-1}). All models except SL-AV and H-AROME have median wind speeds for cloudy conditions larger than observations; median values being 5.5 ms^{-1} for 10-m observations, 6 ms^{-1} for CAFS, 7.5 ms^{-1} for ARPEGE, 6.5 ms^{-1} for DWD, 7.5 ms^{-1} for IFS, and 8 ms^{-1} for NAVY. These differences indicate more relatively higher wind speeds in the models during cloudy skies. For example, ARPEGE has 20% more wind speeds greater than 7.5 ms^{-1} than observations. H-AROME median values are essentially equal to observations and SL-AV median wind speed is 4.5 ms^{-1}.

4.6. Near-surface stratification biases

Figure 7 shows PDFs for near-surface thermal stratification for all-sky, clear-sky, and cloudy distributions. The first thing to note is that the observed distribution is flatter

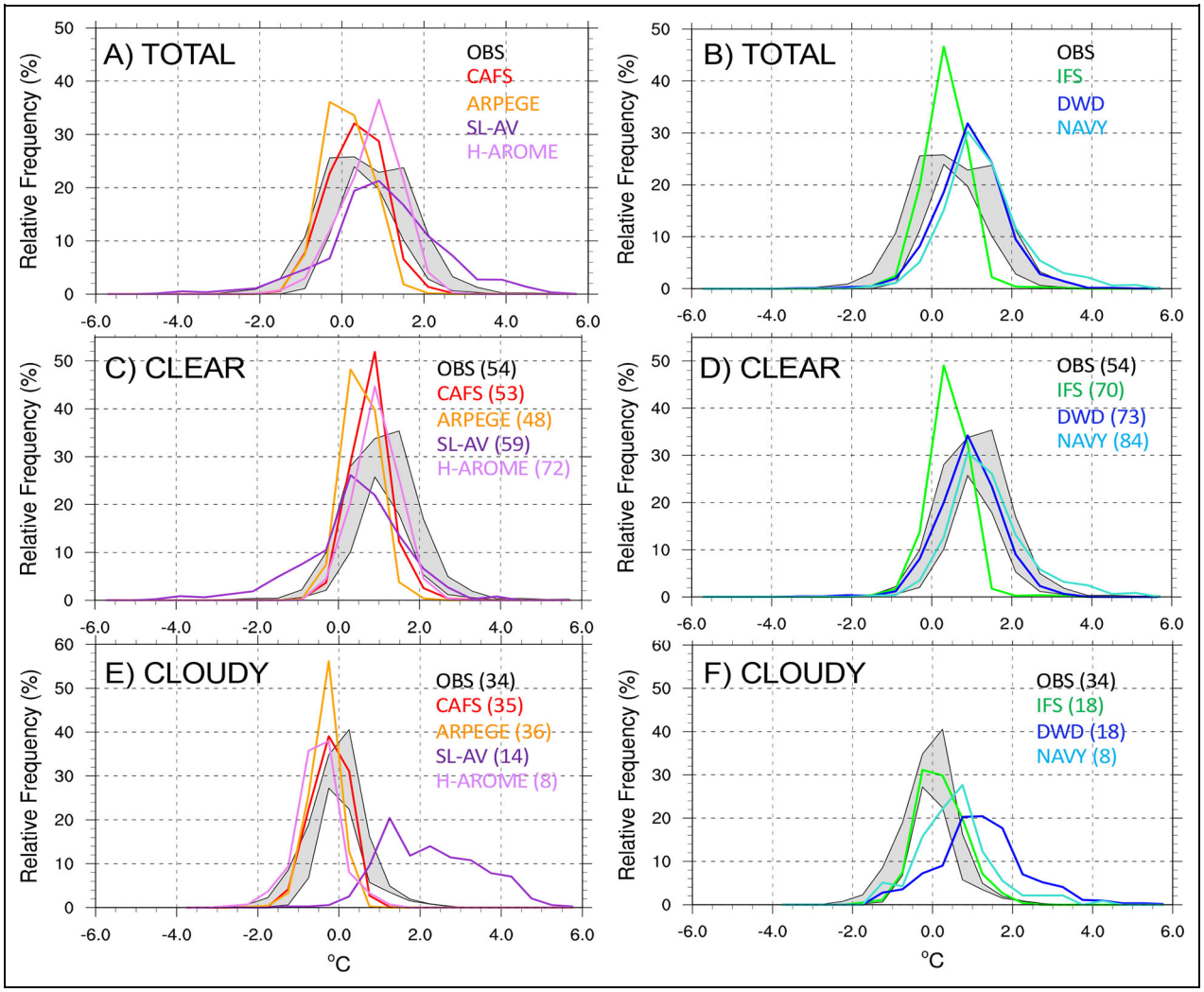

Figure 7. Probability Distribution Functions (PDFs) of model hourly averaged 2-m minus surface temperature difference, in units of °C, using 1 h to 2 day lead times. (A and B) Using all hourly samples. (C and D) Using hourly clear-sky samples. (E and F) Using hourly cloudy samples. The gray shading shows the range of the observed distributions using hourly averaged measurements from the 4 MOSAiC sites. Individual model distributions shown with colored lines as indicated in the legends. The numbers in the legends for the clear and cloudy distributions show the percent of the total sample used in the PDFs.

than the models. This is because of the separation in the PDFs for clear-sky (**Figure 7C** and **D**) and cloudy (**Figure 7E** and **F**) occurrences. It is challenging to reduce the findings from this figure to general conclusions but 2 issues stand out. The first is that all models in (A) and the IFS in (B) have clear-sky distributions that peak at smaller stratification. All of these models except IFS also have cloudy distributions with more negative stratification occurrences than observations, that is, both clear-sky and cloudy distributions are shifted toward less stable conditions. The second issue is that all models except CAFS and ARPEGE dramatically underestimate cloudy occurrences, which causes the all-sky distributions to be more similar to the clear-sky distributions. Other interesting findings to note are that DWD simulates the observed distribution for clear-skies, while NAVY overestimates the occurrence of the very stably stratified conditions. Biases for clear-sky and cloudy conditions tend to compensate for the SL-AV model. It is important to note that 2-m temperature and 10-m winds are derived fields in the models and is not used directly in the simulations but, as is the case for some models such as the CAFS model, is derived using stability functions that are also used in the turbulent heat flux calculations.

4.7. Sensible heat flux biases

Figure 8 shows the surface sensible heat flux PDFs for all-sky, clear-sky, and cloudy distributions from the models and the 4 MOSAiC observational sites. All models show a larger negative skew for the all-sky distributions than observations. This skew is less pronounced for the 2 models that have a bimodal LWNET distribution (CAFS, ARPEGE). All models except SL-AV more closely represent the clear-sky flux but still have heavier negative tails than observations. For cloudy skies, all models in **Figure 8E** closely follow the observed distribution except SL-AV, which has primarily negative sensible heat flux for cloudy conditions. All models in **Figure 8F**, which

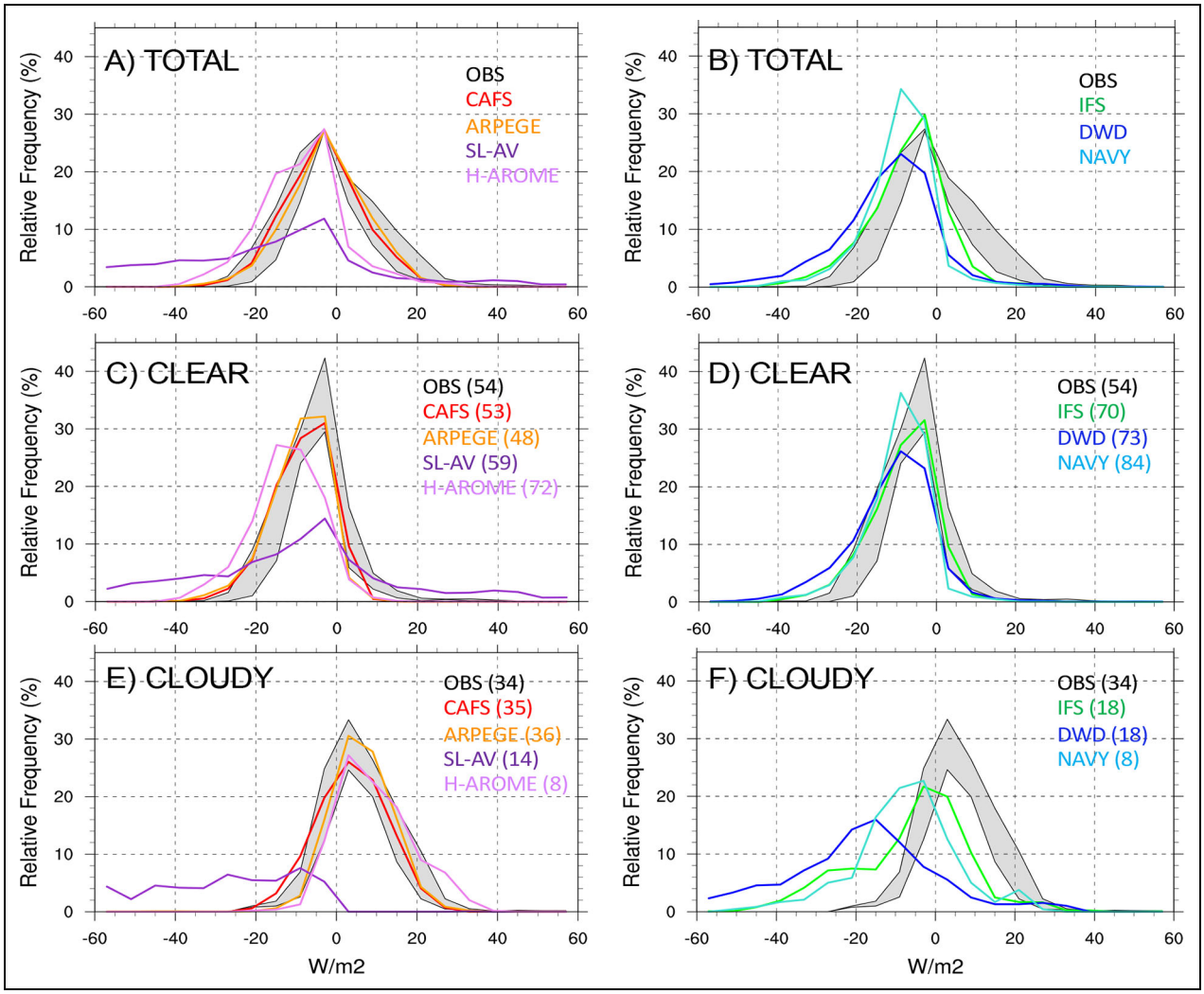

Figure 8. Probability Distribution Functions (PDFs) of model hourly averaged 10-m sensible heat fluxes, in units of Wm^{-2}, using 1 h to 2 day lead times. (A and B) Using all hourly samples. (C and D) Using hourly clear-sky samples. (E and F) Using hourly cloudy samples. The gray shading shows the range of the observed distributions using hourly averaged measurements from the 4 MOSAiC sites. Individual model distributions shown with colored lines as indicated in the legends. The numbers in the legends for the clear and cloudy distributions show the percent of the total sample used in the PDFs.

underestimate the cloudy mode, have negative tails that are larger than for clear-sky conditions, indicating that the downward heat flux for cloudy conditions exceeds the downward heat fluxes for clear-sky conditions. This variability is outside the observations from the MOSAiC campaigns.

4.8. Coupled process relationships

4.8.1. Joint PDFs of LWNET and near-surface stratification

This section focuses on coupled process relationships to identify the processes responsible for model shortcomings and to quantify biases in the simulations of coupled feedbacks. The first relationship is between the near-surface thermal stratification and LWNET (**Figure 9**). Observations show a bimodal distribution with a clear-sky mode, where upward surface longwave radiation dominates over downward longwave radiation causing a cooling of the surface, producing a stably stratified near-surface layer, and a cloudy mode, where downward longwave radiation almost compensates for the upward surface longwave radiation and cloud-driven turbulence causes the near-surface environment to be well-mixed (**Figure 9A**). Observations indicate ranges of variability to assess the models; the cloudy mode peaks with LWNET less than −20 Wm^{-2} and clear-sky mode peaks with LWNET less than −55 Wm^{-2} or greater than −30 Wm^{-2} are outside the observed variability.

Since only CAFS and IFS maintain liquid in clouds at cold temperatures (**Figure 4**), it is interesting to see a cloudy mode in the ARPEGE joint PDF. This is due to ice clouds in ARPEGE that produce a cloudy mode with a peak in LWNET and near-surface thermal stratification similar to the mixed-phase clouds in the observations and the CAFS model. It is very interesting to see that the SL-AV and H-AROME models produce clear-sky modes with LWNET 10–20 Wm^{-2} smaller than observations, presumably due to the colder surface temperatures in H-AROME

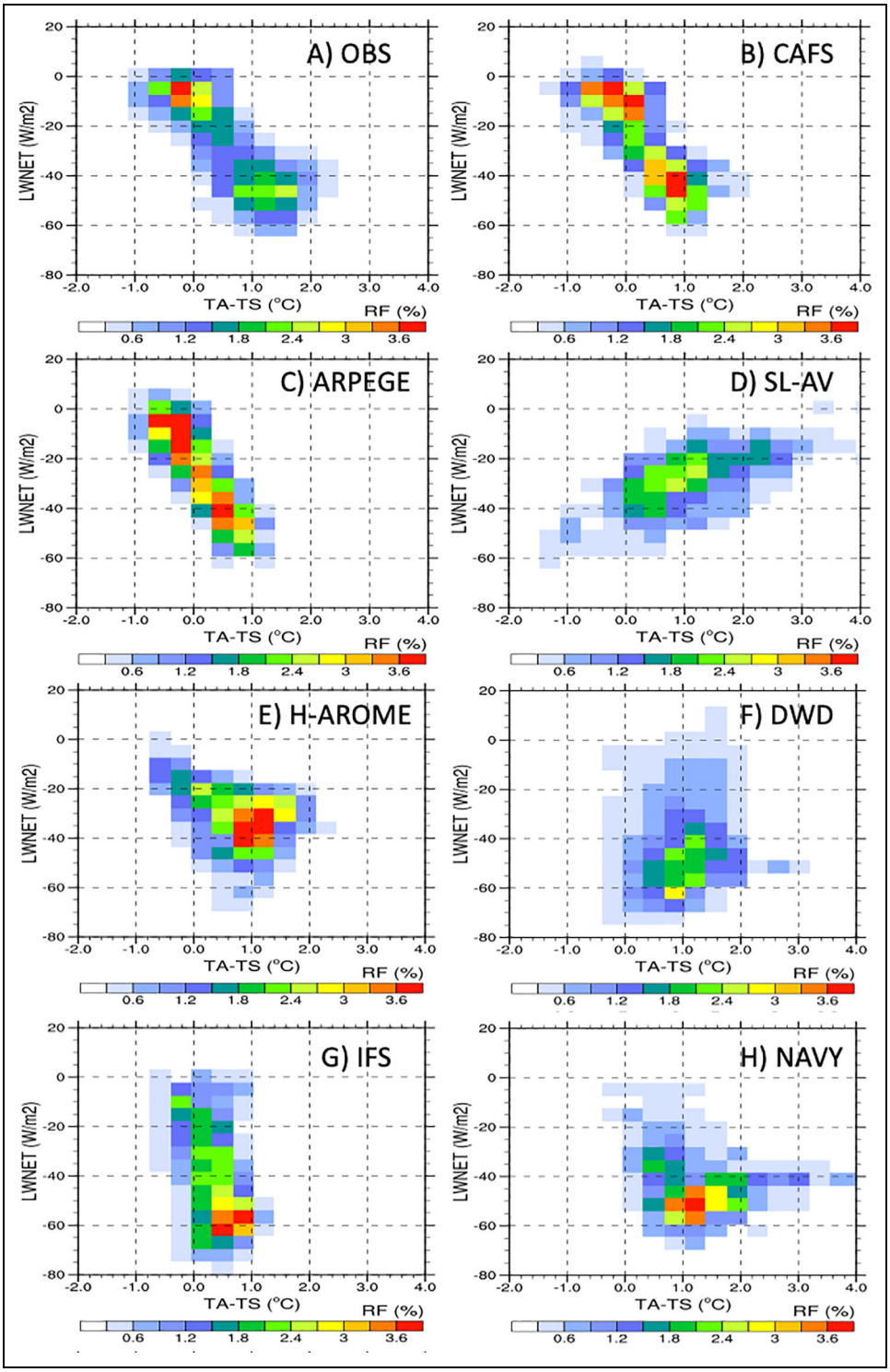

Figure 9. Joint Probability Distribution Functions (PDFs) of modeled wintertime near-surface thermal stratification and net longwave radiation, in units of °C and Wm^{-2}. (A) The observed distribution using hourly averaged measurements from the 4 MOSAiC sites. (B) CAFS, (C) ARPEGE, (D) SL-AV, (E) H-AROME, (F) IFS, (G) DWD, (H) NAVY.

(**Figure 4**) and the larger downward longwave flux in SL-AV (**Figure 2**). But it should be noted that this does not correspond to a shift in the clear-sky mode near-surface thermal stratification, potentially due to the parameterizations that maintain the near-surface thermal stratification (e.g., **Figure 8**). This relationship is also of interest for the 2 models without a distinct cloudy mode (DWD, NAVY) and IFS, which produces cloud liquid with small amounts and underestimates the cloudy mode. There is an indication that DWD and NAVY have more cases with $\Delta T > 2°C$ than observations but the peak in the clear-sky near-surface temperature difference is within the observed range (0.9 C–1.6°C). The clear-sky values for IFS are smaller than the observed range.

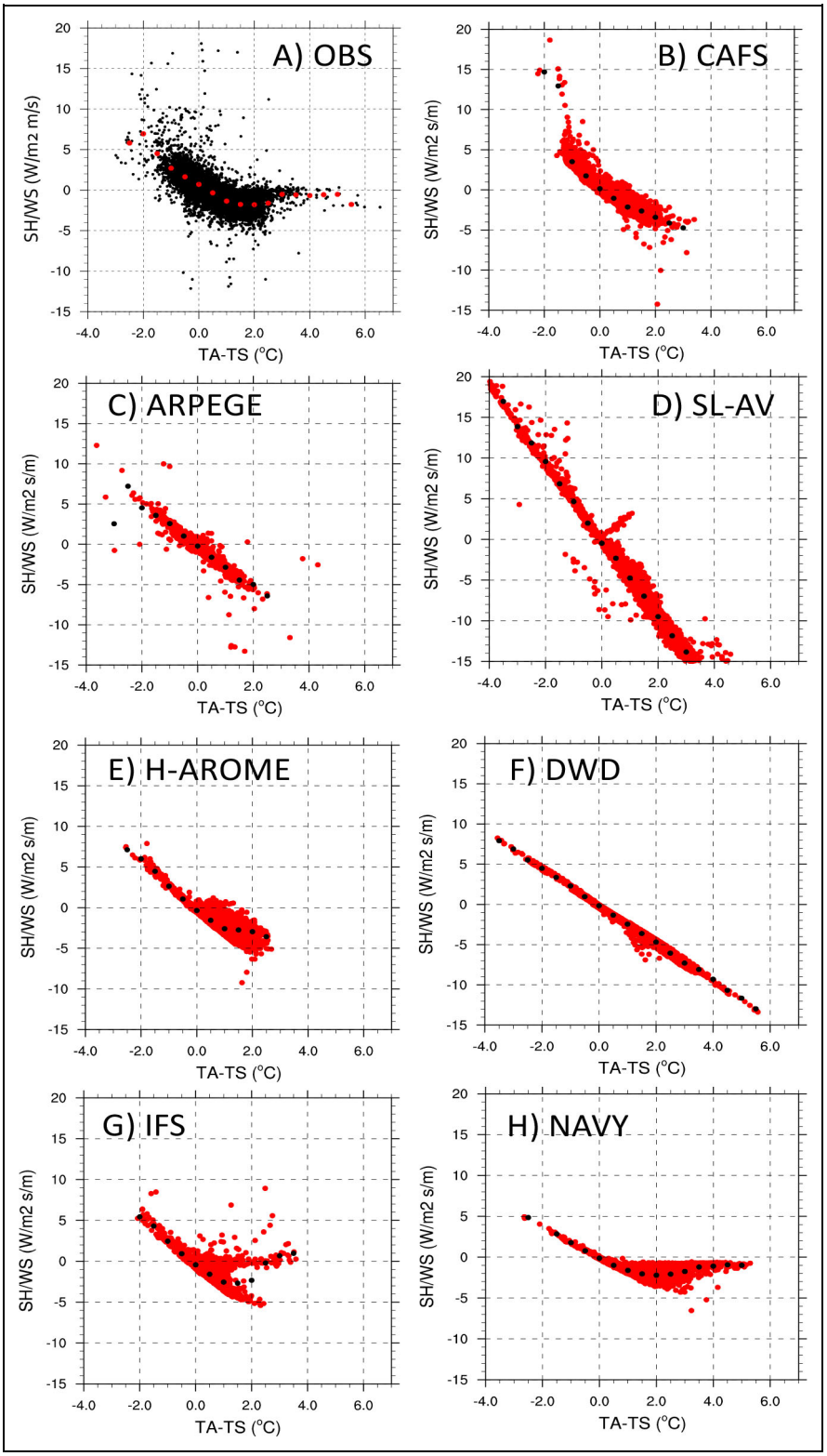

Figure 10. Scatterplot of hourly near-surface thermal stratification versus scaled sensible heat fluxes, in units of °C and $Wm^{-3}s$. (Black) red dots show the individual samples, (red) black dots show the 0.5°C binned values in observations (models). (A) OBS, (B) CAFS, (C) ARPEGE, (D) SL-AV, (E) H-AROM, (F) IFS, (G) DWD, (H) NAVY.

4.8.2. Scaled sensible heat flux versus near-surface stratification

To get insight into biases due to the sensible heat flux parameterizations used in the models, **Figure 10** shows scatterplots of the scaled sensible heat flux (sensible heat flux divided by 10-m wind speed) relative to the near-surface temperature difference. The slope of this relationship is the diagnosed transfer coefficient in the parameterization for the sensible heat flux (see Tjernstroüm et al., 2005). The red dots in **Figure 10A** show the binned values using surface sensible heat flux and temperature and 10-m wind speed from the 4 MOSAiC sites.

The first thing to note in **Figure 10** is that all models have a tighter spread than the observations, particularly for near neutral stratification. This indicates that the observed scaled sensible heat flux is not strictly a function of the near-surface thermal stratification, potentially due to transitions, that is, when the gradient and the flux do not balance. Observations show a reduction in the scaled heat fluxes for increasing near-surface thermal stratification ($\Delta T > 1.5^oC$). Four of the models simulate this variability; CAFS, H-AROME, IFS, and NAVY. Only 2 of the models produce the decline of the scaled sensible heat flux observed for strongly stably stratified near-surface conditions (IFS and NAVY). However, IFS has a bifurcation with a reduction in the scaled sensible heat flux for only a fraction of the occurrences with $\Delta T > 1^oC$. The 3 other models have a constant slope for this relationship, with ARPEGE being limited to $\Delta T < 2^oC$, while SL-AV and DWD produce a constant slope out to $\Delta T > 3^oC$. This is unrealistic because observations indicate diminishing scaled sensible heat flux as thermal stratification continues to increase. It is not clear what is limiting the occurrence of the strongly stable stratification cases in CAFS, ARPEGE, and H-AROME. This may be due to sensible heat fluxes that are too large as is seen in CAFS and ARPEGE but H-AROME has scaled sensible heat fluxes close to the observations for $\Delta T = 2^oC$. Another factor may be sea ice concentration less than 99.5% in these models. However, SL-AV and IFS also have sea ice concentrations less than 99.5% and these models simulate occurrences with $\Delta T > 3^oC$. This issue will be investigated in a follow-up study.

It is also interesting to note that CAFS is the only model that has relatively few unstable cases with $\Delta T < -1.5^oC$. Since this is the only model with cloud liquid water similar to observations, it was anticipated that the thermal stratification during cloudy conditions would also be similar to observations. It is not clear how the models with very little cloud liquid and ice can produce unstably stratified near surface conditions during the winter season. The follow-up study will focus on looking more closely at the sensible heat flux parameterizations used in these models in a specifically designed testbed.

Figure 11 shows scatterplots of the 3 terms in the surface energy budget (Equation 2); LWNET shown in colors separated into cloudy (black), thin clouds (green), and clear-sky (red) regimes; the sensible heat flux on the y-axis; the conductive flux calculated as a residual on the x-axis. This diagnostic is designed to identify whether the models are reproducing the observed relationship between these 3 terms and is not meant to display the frequency of occurrences. Note the ranges used for SL-AV and DWD differ from the other scatterplots.

Considering the cloudy occurrences first, the observed scatter for cloudy occurrences falls on the 1-to-1 line for positive (upward) sensible heat flux and conductive flux due to small LWNET. It is seen that there are only a few observed occurrences for negative (downward) sensible heat flux that coincides with negative (downward) conductive flux, whereas 4 model have frequent occurrences in this regime, especially the 3 models that underestimate the cloudy mode (DWD, IFS, NAVY) and SL-AV. The more frequent occurrence of variability in the models in this regime and the values with magnitudes larger than 20 Wm^{-2} are outside the observed results. H-AROME has scatter for cloudy occurrences outside the 1-to-1 line, where upward conductive flux can be over 2 times the upward sensible heat flux.

For the clear-sky occurrences, CAFS, ARPEGE, and H-AROME generally reproduce the observed variability. However, all 3 models produce only a limited number of cases when sensible heat flux is positive, whereas observations have sensible heat flux reaching values of up to 60 Wm^{-2}. There is significant scatter outside the observed range for clear-skies for the 3 models that underestimate the cloudy mode. All 3 models produce upward conductive flux greater than 60 Wm^{-2}, as well as downward sensible heat flux with magnitudes greater than 30 Wm^{-2}. The indication is that a persistent clear-sky regime at times pushes the model into an unrealistic surface energy budget or the models are tuned to compensate for the underestimate of the cloudy mode. The SL-AV model also has similar biases in clear-sky occurrences, but over a much wider range. There is also interesting observed sensible heat flux scatter for clear-sky and cloudy sky conditions that is greater than 40 Wm^{-2}, which is only reproduced in SL-AV.

5. Summary and discussion

This study evaluates the representation of wintertime statistics of the atmospheric boundary layer and surface energy budget in the Central Arctic in weather-scale operational and experimental forecasts systems with observations taken during the MOSAiC campaign. Statistical distributions and process diagnostics are used to identify systematic biases against observations of near-surface atmospheric structure (from 3 separate atmospheric flux stations and a 10-m tower located 10–20 km apart), cloud characteristics (based on ground-based remote sensors on the *Polarstern*), and surface fluxes (from 3 separate atmospheric flux stations and the Met City installation). Forecasts from 7 experimental and operational forecast systems are included in the intercomparison. The model domains of the studied systems range from global domains to regional pan-Arctic or Central Arctic domains. All models have prognostic surface temperature but sea ice and snow are treated with different levels of complexity across the systems, ranging from dynamic sea ice models with multilevel snow models to fixed sea ice thickness and concentration with no snow on sea ice. The majority of the systems are fully coupled ocean-sea ice-atmosphere models but 2 systems included in this intercomparison have fixed sea ice. Differences in the forecast system configurations provide an opportunity to identify systematic biases and impacts on the near-surface atmosphere and surface energy budget.

Individual and systematic findings from this multimodel intercomparison using MOSAiC observations are summarized in **Table 4**. This study provides a benchmark of the performance in the Arctic of current operational systems, which can be revised and further developed informed by this new observational dataset and the identified deficiencies. The models in general are not able to

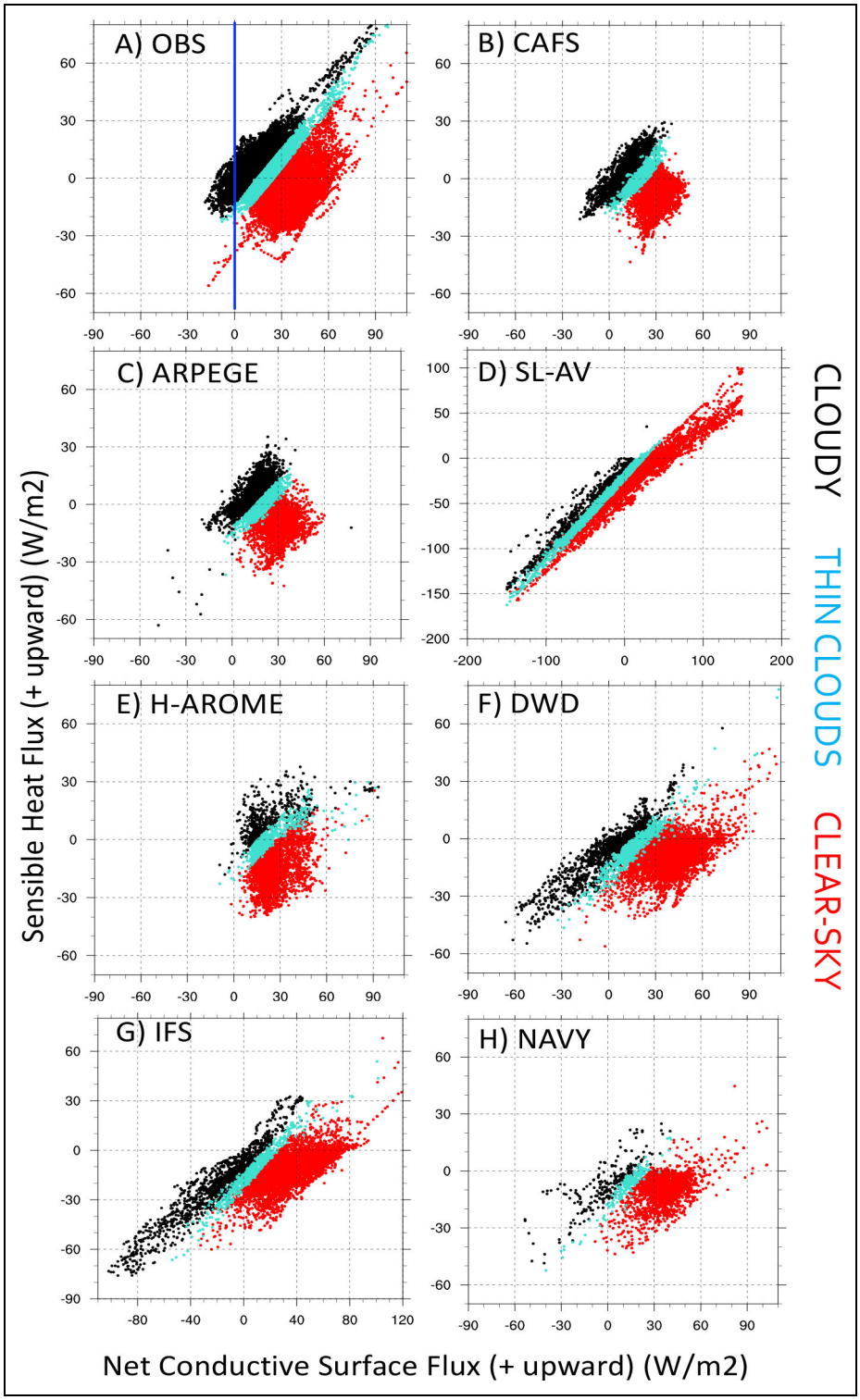

Figure 11. Scatterplot of net conductive surface flux versus sensible heat flux for cloudy skies (black), thin clouds (turquoise), and clear-sky (red), in units of Wm^{-2}. (A) OBS, (B) CAFS, (C) ARPEGE, (D) SL-AV, (E) H-AROME, (F) IFS, (G) DWD, (H) NAVY.

reproduce the observed bimodal LWNET distribution with the exception of 2 models. Models generally struggle to represent thin liquid clouds in the Arctic which impacts the LWD. This bias is found in one model to be compensated by a positive biased distribution of surface temperatures that yields a LWNET distribution closer to the observed. All models show, as expected, a relationship between the scaled sensible heat flux and the near surface stability. However, they have less spread and cover a different parameter range, especially for highly stably stratified conditions, than the observations, indicating that the observed scaled sensible heat flux is not strictly a function of the near-surface thermal stratification. About half of the models produce too many occurrences in the regime

Table 4. A summary of systematic and individual model (Table 3) issues from the intercomparison informed by observations

Major Findings	Summary
Simulating observed bimodal LWNET	Only 2 models have a bimodal LWNET distribution similar to observations (CAFS and ARPEGE) and 2 models have a unimodal distribution with a peak between the observed clear-sky and cloudy modes (SL-AV and H-AROME), all 4 of these models produce distributions with lighter tails than observations, indicating fewer extreme values. Three models (IFS, DWD, NAVY) significantly underestimate the cloudy mode and have clear-sky distributions shifted toward larger LWNET magnitudes.
Compensating biases in LWNET	Even though CAFS and ARPEGE have similar LWNET distributions, ARPEGE overestimates the fractional occurrence of thin clouds, underestimates the occurrence of clear-skies, and simulates a surface temperature distribution that essentially removes the observed positive skew toward warm temperatures and shifts the modal peak by 5.5°C.
Biases in LWNET due to LWU	The SL-AV and H-AROME biases in LWNET can be explained as the absence of a cloudy mode and a 20 Wm^{-2} decrease in magnitude of the LWU mode peak, primarily due to a shift in the clear-sky occurrences in H-AROME and a shift in the cloudy occurrences for SL-AV. Cloudy occurrences have colder temperatures than clear-sky occurrences in SL-AV and are the primary cause for the shift toward colder temperatures in the total distribution.
Biases in LWNET due to LWD	IFS, DWD, and NAVY show a shift in the clear-sky mode toward larger negative values relative to observations. The NAVY bias is primarily due to clear-sky LWD being too small, while DWD and IFS have clear-sky LWD that is too large but this is compensated by surface temperature that is too warm. All 3 models underestimate the cloudy mode but have different surface temperature distributions, with NAVY median temperatures being 4°C too low, IFS median temperatures being 2°C too high, and DWD median temperatures being very close to observations.
Simulating observed LWP	Only CAFS produces LWP similar to observations. IFS produces limited cloud liquid water, but underestimates the cloudy mode. All models produce cloud ice, but only ARPEGE produces enough cloud ice to create a LWNET similar to observations and CAFS even though this model does not maintain liquid in clouds at the coldest temperatures.
Compensating biases due to SH parameterizations	SL-AV and H-AROME produce clear-sky modes with the magnitude of LWNET 10–20 Wm^{-2} smaller than observations, presumably due to the colder surface temperatures in H-AROME and the larger downward longwave flux in SL-AV. But it is interesting to note that this does not correspond to a shift in the clear-sky mode near-surface thermal stratification, potentially due to the parameterizations that maintains the near-surface thermal stratification.
Relationship between scaled sensible heat flux and near-surface thermal stratification	All models have a tighter spread in the scatter between scaled sensible heat flux and near-surface thermal stratification than the observations indicating that the observed scaled sensible heat flux is not strictly a function of the near-surface thermal stratification, that is, includes transitions when the gradient and flux do not balance. SL-AV and DWD have a constant slope for this relationship and have more occurrences of near-surface temperature differences greater than 4°C, which is counter-intuitive since this means the parameterizations for these 2 models continue to produce larger sensible heat flux as thermal stratification increases. IFS and NAVY are the only models that produce the decline of the scaled sensible heat flux for strongly stable near-surface conditions. All models except H-AROME have frequent occurrences in an unobserved regime for cloudy skies with downward sensible heat flux and downward conductive flux.
Relationship between SEB terms	All models except H-AROME produce too many occurrences in the regime with downward sensible heat flux and downward conductive flux. For the 3 models that underestimate the cloudy mode, a persistent clear-sky regime at times pushes the model into unrealistic surface energy budgets. One unobserved regime for clear skies is large downward sensible heat flux (magnitudes greater than 30 Wm^{-2}) with conductive flux close to zero. The other unobserved regime for clear skies is large upward conductive flux (greater than 60 Wm^{-2}) with sensible heat flux close to zero. H-AROME has cloudy scatter outside the 1-to-1 line, where upward conductive flux is over 2 times the upward sensible heat flux.

The acronyms include net longwave radiation at the surface (LWNET), upward longwave flux at the surface (LWU), downward longwave flux at the surface (LWD), liquid and ice water paths (LWP and IWP), and surface sensible heat flux (SH).

with downward sensible heat flux and downward conductive flux.

The wintertime Arctic SEB is a relatively simple balance between net longwave radiation, sensible heat flux, and conductive flux. Therefore, a bias in one of these components manifests as a compensating bias in at least one of the others, that is, the compensation is not necessarily a due to unrealistic model physics in another component. In this study we see many different manifestations of model bias and the potential implications of those biases on multiple terms. Due to the connected nature of these terms, it is also difficult to fully diagnose the root causes of model deficiencies, since the compensation sometimes can mask those causes. Based on the analysis presented, there appears to be 3 general types of challenges within the models: representing the radiative impact of clouds, representing the interaction of atmospheric heat fluxes with subsurface fluxes (i.e., snow and ice properties), and representing the relationship between the stability regime and turbulent heat fluxes.

The SEB is also sensitively dependent on the representation of snow over sea ice. Snow on sea ice is an insulator that limits the warming of the atmosphere by the underlying ocean and limits the growth of sea ice by the loss of surface energy. Snow on sea ice is represented in the CAFS and ARPEGE models with 1 layer, in the H-AROME model by 12 layers, and neglected in the other 4 models. The lack of snow on sea ice in DWD and IFS likely explains a significant part of the warm bias in surface temperature during clear-sky conditions. Indeed, a recent study demonstrates a positive impact on boundary layer structure and surface temperature in a version of the IFS with snow on sea ice, but also highlights that improvements in surface physics need to be matched by improvements in the representation of Arctic clouds (Arduini et al., 2022). Clearly, the representation of snow, depth and characteristics, on sea ice is a critically important part of the coupled system, the model information collected for this study is not enough to isolate its impact on the surface energy balance within different model configurations. A follow-up study will focus on looking more closely at parameterizations and configurations used in the models in a testbed with MOSAiC observations designed to produce more detailed and comprehensive comparisons.

Observations from MOSAiC, SHEBA, and other Central Arctic campaigns provide the opportunity to evaluate and improve the representation of coupled processes unique to the Arctic. Clearly, state-of-the art forecast systems do not adequately simulate the Arctic system in the Central Arctic during winter. It is striking how each model assessed here has a distinct set of shortcomings. Improved parameterizations of cloud processes and boundary layer turbulence are needed, as well as, improved initial conditions and the representation of snow on sea ice, among other model configurations. Targeted model studies are required to make progress and to improve forecasts of the Arctic system and projections of the role of the Arctic in the climate system.

Data accessibility statement

All observations used in this study are publicly available on the MOSAiC archives, including PANGAEA (https://www.pangaea.de/), the Department of Energy Atmospheric Radiation Measurement (ARM) Program data archive (https://adc.arm.gov/discovery/), and the National Science Foundation's Arctic Data Center (http://arcticdata.io). All model output is available on the YOPP Data Portal (https://yopp.met.no/).

Acknowledgments

Data used in this manuscript were produced as part of the international Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) with the tag MOSAiC20192020. We thank all persons involved in the expedition of the Research Vessel *Polarstern* during MOSAiC in 2019–2020 (AWI_PS122_00) as listed in Nixdorf et al. (2021). We thank the WWRP Year of Polar Prediction Processes Task Team for productive and insightful discussions. A subset of data was obtained from the Atmospheric Radiation Measurement (ARM) User Facility, a U.S. Department of Energy (DOE) Office of Science User Facility Managed by the Biological and Environmental Research Program. We thank Rostislav Fadeev (INM RAS) for modifying the SL-AV model code to produce output at the MOSAiC location. NPB and JDD acknowledge the support of the Office of Naval Research PE 0601153 N and the contributions of Andrew Huang for post processing the Navy ESPC data. The authors thank the anonymous reviewers; their constructive comments have helped to improve the manuscript.

Funding

- US National Science Foundation Office of Polar Programs (OPP-1724551)
- US Department of Energy—Office of Science Atmospheric Radiation Measurement and Atmospheric System Research Programs (DE-SC0021341)
- US National Oceanic and Atmospheric Administration Physical Sciences Laboratory and Arctic Research Program
- Russian Science Foundation (21-17-00254)
- Norwegian Research Council project no. 280573 "Advanced models and weather prediction in the Arctic: enhanced capacity from observations and polar process representations (ALERTNESS)"
- J Day was supported by the European Union funded project INTERACT 3 (GA 871120)

Competing interests

All authors declare that they have no competing interests. MDS is a guest editor for the Elementa Special Feature on MOSAiC but was not involved in the editorial or review process for this manuscript.

Author contributions

Contributed to conception and design: AS, GS, JJD, MDS.

Contributed to acquisition/production of data: AS, MDS, JJD, NPB, EB, HPF, SK, YB, TR, MT.

Contributed to analysis and interpretation of data: All authors.

Drafted and/or revised the manuscript: All authors.

Approved and submitted manuscript: All authors.

Submitted the manuscript: AS.

References

Arduini, G, **Keeley, S**, **Day, JJ**, **Sandu, I**, **Zampieri, L**, **Balsamo, G.** 2022. On the importance of representing snow over sea-ice for simulating the Arctic boundary layer. *Journal of Advances in Modeling Earth Systems* **14**: e2021MS002777. DOI: http://dx.doi.org/10.1029/2021MS002777.

Barton, N, **Metzger, EJ**, **Reynolds, CA**, **Ruston, B**, **Rowley, C**, **Smedstad, OM**, **Ridout, JA**, **Wallcraft, A**, **Frolov, S**, **Hogan, P**, **Janiga, MA**, **Shriver, JF**, **McLay, J**, **Thoppil, P**, **Huang, A**, **Crawford, W**, **Whitcomb, T**, **Bishop, CH**, **Zamudio, L**, **Phelps, M.** 2021. The Navy's Earth System Prediction Capability: A new global coupled atmosphere–ocean–sea ice prediction system designed for daily to subseasonal forecasting. *Earth and Space Science* **8**(4): e2020EA001199. DOI: http://dx.doi.org/10.1029/2020EA001199.

Batrak, Y, **Kourzeneva, E**, **Homleid, M.** 2018. Implementation of a simple thermodynamic sea ice scheme, SICE version 1.0-38h1, within the ALADIN–HIRLAM numerical weather prediction system version 38h1. *Geoscientific Model Development* **11**: 3347–3368. DOI: http://dx.doi.org/10.5194/gmd-11-3347-2018.

Batrak, Y, **Müller, M.** 2019. On the warm bias in atmospheric reanalyses induced by the missing snow over Arctic sea-ice. *Nature Communications* **10**: 4170. DOI: http://dx.doi.org/10.1038/s41467-019-11975-3.

Bazile, E, **Azouz, N**, **Napoly, A**, **Loo, C.** 2020. Impact of the 1D sea-ice model GELATO in the global model ARPEGE. Available at http://bluebook.meteoinfo.ru/index.php?year=2020&ch_=2.

Beesley, JA, **Bretherton, CS**, **Jakob, C**, **Andreas, EL**, **Intrieri, JM**, **Uttal, TA**. 2000. A comparison of cloud and boundary layer variables in the ECMWF forecast model with observations at Surface Heat Budget of the Arctic Ocean (SHEBA) ice camp. *Journal of Geophysical Research* **105**: 12337–12349. DOI: http://dx.doi.org/10.1029/2000JD900079.

Bengtsson, L, **Andrae, U**, **Aspelien, T**, **Batrak, Y**, **Calvo, J**, **de Rooy, W**, **Gleeson, E**, **Hansen-Sass, B**, **Homleid, M**, **Hortal, M**, **Ivarsson, K-I**, **Lenderink, G**, **Niemelä, S**, **Nielsen, KP**, **Onvlee, J**, **Rontu, L**, **Samuelsson, P**, **Muñoz, DS**, **Subias, A**, **Tijm, S**, **Toll, V**, **Yang, X**, **Køltzow, MO.** 2017. The HARMONIE–AROME model configuration in the ALADIN–HIRLAM NWP system. *Monthly Weather Review* **145**: 1919–1935. DOI: http://dx.doi.org/10.1175/MWR-D-16-0417.1.

Brunke, MA, **Zhou, M**, **Zeng, X**, **Andreas, EL.** 2006. An intercomparison of bulk aerodynamic algorithms used over sea ice with data from the Surface Heat Budget for the Arctic Ocean (SHEBA) experiment. *Journal of Geophysical Research* **111**: C09001. DOI: http://dx.doi.org/10.1029/2005JC002907.

Cox, C, **Gallagher, M**, **Shupe, M**, **Persson, O**, **Solomon, A**, **Ayers, T**, **Costa, D**, **Hutchings, J**, **Leach, J**, **Morris, S**, **Osborn, J**, **Pezoa, S**, **Uttal, T.** 2021a. Atmospheric Surface Flux Station #30 measurements (Level 1 Raw), Multidisciplinary Drifting Observatory for the Study of Arctic Climate (MOSAiC), central Arctic, October 2019–September 2020. Arctic Data Center. DOI: http://dx.doi.org/10.18739/A20C4SM1J. Accessed January 18, 2023.

Cox, C, **Gallagher, M**, **Shupe, M**, **Persson, O**, **Solomon, A**, **Ayers, T**, **Costa, D**, **Hutchings, J**, **Leach, J**, **Morris, S**, **Osborn, J**, **Pezoa, S**, **Uttal, T.** 2021b. Atmospheric Surface Flux Station #40 measurements (Level 1 Raw), Multidisciplinary Drifting Observatory for the Study of Arctic Climate (MOSAiC), central Arctic, October 2019–September 2020. Arctic Data Center. DOI: http://dx.doi.org/10.18739/A2CJ87M7G. Accessed January 18, 2023.

Cox, C, **Gallagher, M**, **Shupe, M**, **Persson, O**, **Solomon, A**, **Ayers, T**, **Costa, D**, **Hutchings, J**, **Leach, J**, **Morris, S**, **Osborn, J**, **Pezoa, S**, **Uttal, T**. 2021c. Atmospheric Surface Flux Station #50 measurements (Level 1 Raw), Multidisciplinary Drifting Observatory for the Study of Arctic Climate (MOSAiC), central Arctic, October 2019–September 2020. Arctic Data Center. DOI: http://dx.doi.org/10.18739/A2445HD46. Accessed January 18, 2023.

Cox, C, **Gallagher, M**, **Shupe, M**, **Persson, O**, **Solomon, A**, **Ayers, T**, **Costa, D**, **Hutchings, J**, **Leach, J**, **Morris, S**, **Osborn, J**, **Pezoa, S**, **Uttal, T.** 2021d. 10-meter (m) meteorological flux tower measurements (Level 1 Raw), Multidisciplinary Drifting Observatory for the Study of Arctic Climate (MOSAiC), central Arctic, October 2019–September 2020. Arctic Data Center. DOI: http://dx.doi.org/10.18739/A2VM42Z5F. Accessed January 18, 2023.

Haiden, T, **Janousek, M**, **Vitart, F**, **Ferranti, L**, **Prates, F.** 2019. Evaluation of ECMWF forecasts, including the 2019 upgrade. DOI: http://dx.doi.org/10.21957/mlvapkke.

Hunke, EC, **Lipscomb, WH**. 2008. CICE: The Los Alamos Sea Ice Model. Documentation and Software User's Manual, Version 4.0. T-3 Fluid Dynamics Group, Los Alamos National Laboratory. Technical Report. LA-CC-06-012.

Inoue, J, **Sato, K**, **Rinke, A**, **Cassano, JJ**, **Fettweis, X**, **Heinemann, G**, **Matthes, H**, **Orr, A**, **Phillips, T**, **Seefeldt, M**, **Solomon, A**, **Webster, S.** 2020. Clouds and radiation processes in regional climate models evaluated using observations over the ice-free Arctic Ocean. *Journal of Geophysical Research* **126**. DOI: http://dx.doi.org/10.1029/2020JD033904.

Jordan, RE, **Andreas, EL**, **Makshtas, AP.** 1999. Heat budget of snow-covered sea ice at North Pole 4. *Journal*

of Geophysical Research: Oceans **104**: 7785–7806. DOI: http://dx.doi.org/10.1029/1999JC900011.

Kauffman, BG, **Large, WG**. 2002. The CCSM coupler, version 5.0.1. Available at https://github.com/CICE-Consortium/CICE/blob/master/doc/PDF/KL_NCAR2002.pdf.

Keeley, S, **Mogensen, K.** 2018. Dynamic sea ice in the IFS, ECMWF Newsletter. DOI: http://dx.doi.org/10.21957/4ska25furb.

Knust, R. 2017. Polar research and supply vessel POLARSTERN operated by the Alfred-Wegener-Institute. *Journal of Large-Scale Research Facilities JLSRF* **3**. DOI: http://dx.doi.org/10.17815/jlsrf-3-163.

Køltzow, M, **Casati, B**, **Bazile, E**, **Haiden, T**, **Valkonen, T.** 2019. An NWP model intercomparison of surface weather parameters in the European Arctic during the year of polar prediction special observing period Northern Hemisphere 1. *Weather and Forecasting* **34**(4): 959–983. DOI: https://doi.org/10.1175/WAF-D-19-0003.1.

Kondo, J, **Kanechika, O**, **Yasuda, N.** 1978. Heat and momentum transfers under strong stability in the atmospheric surface layer. *Journal of the Atmospheric Sciences* **35**: 1012–1021. DOI: http://dx.doi.org/10.1175/1520-0469(1978)035<1012:HAMTUS>2.0.CO;2.

Krumpen, T, **Birrien, F**, **Kauker, F**, **Rackow, T**, **von Albedyll, L**, **Angelopoulos, M**, **Belter, HJ**, **Bessonov, V**, **Damm, E**, **Dethloff, K**, **Haapala, J**, **Haas, C**, **Harris, C**, **Hendricks, S**, **Hoelemann, J**, **Hoppmann, M**, **Kaleschke, L**, **Karcher, M**, **Kolabutin, N**, **Lei, R**, **Lenz, J**, **Morgenstern, A**, **Nicolaus, M**, **Nixdorf, U**, **Petrovsky, T**, **Rabe, B**, **Rabenstein, L**, **Rex, M**, **Ricker, R**, **Rohde, J**, **Shimanchuk, E**, **Singha, S**, **Smolyanitsky, V**, **Sokolov, V**, **Stanton, T**, **Timofeeva, A**, **Tsamados, M**, **Watkins, D**. 2020. The MOSAiC ice floe: Sediment-laden survivor from the Siberian Shelf. *The Cryosphere*. DOI: http://dx.doi.org/10.5194/tc-2020-64.

Lawrence, ZD, **Perlwitz, J**, **Butler, AH**, **Manney, GL**, **Newman, PA**, **Lee, SH**, **Nash, ER**. 2020. The remarkably strong Arctic stratospheric polar vortex of winter 2020: Links to record-breaking Arctic oscillation and ozone loss. *Journal of Geophysical Research* **125**: e2020JD033271. DOI: http://dx.doi.org/10.1029/2020JD033271.

Mironov, D, **Ritter, B**, **Schulz, J-P**, **Buchhold, M**, **Lange, M**, **Machulskaya, E.** 2012. Parameterisation of sea and lake ice in numerical weather prediction models of the German weather service. *Tellus A* **64**(0). DOI: http://dx.doi.org/10.3402/tellusa.v64i0.17330.

Monin, AS, **Obukhov, AM.** 1954. Basic laws of turbulent mixing in the atmosphere near the ground. *Trudy Geofiz Inst AN SSSR* **24**: 163–187.

Moritz, RE, **Curry, JA**, **Thorndike, AS**, **Untersteiner, N.** 1993. SHEBA, a Research Program on the Surface Heat Budget of the Arctic Ocean, Arctic System Science Center: Ocean-Atmosphere Interactions, Seattle, Washington: University of Washington. Rep. No. 3, 34 pp.

Perovich, DK, **Andreas, EL**, **Curry, JA**, **Eiken, H**, **Fairall, CW**, **Grenfell, TC**, **Guest, PS**, **Intrieri, J**, **Kadko, D**, **Lindsay, RW**, **McPhee, MG**, **Morison, J**, **Moritz, RE**, **Paulson, CA**, **Pegau, WS**, **Persson, POG**, **Pinkel, R**, **Richter-Menge, JA**, **Stanton, T**, **Stern, H**, **Sturm, M**, **Tucker III, WB**, **Uttal, T**. 1999. Year on ice gives climate insights. *Eos Transactions* **80**: 485–486. DOI: http://dx.doi.org/10.1029/EO080i041p00481-01.

Perovich, DK, **Grenfell, TC**, **Richter-Menge, JA**, **Light, B**, **Tucker III, WB**, **Eicken, H**. 2003. Thin and thinner: Sea ice mass balance measurements during SHEBA. *Journal of Geophysical Research* **108**: 8050. DOI: http://dx.doi.org/10.1029/2001JC001079.

Persson, O, **Fairall, C**, **Andreas, E**, **Guest, P**, **Perovich, D.** 2002. Measurements near the atmospheric surface flux group tower at SHEBA: Near-surface conditions and surface energy budget. *Journal of Geophysical Research* **107**: 8045. DOI: http://dx.doi.org/10.1029/2000JC000705.

Riihimaki, L. 2021. Radiation Instruments on Ice (ICERADRIIHIMAKI). Atmospheric Radiation Measurement (ARM) user facility. DOI: http://dx.doi.org/10.5439/1608608. Accessed July 15, 2022.

Rinke, A, **Cassano, J**, **Cassano, E**, **Jaiser, R**, **Handorf, D.** 2021. Meteorological conditions during the MOSAiC expedition: Normal or anomalous? *Elementa: Science of the Anthropocene* **9**(1). DOI: http://dx.doi.org/10.1525/elementa.2021.00023.

Rinke, A, **Dethloff, K**, **Cassano, J**, **Christensen, JH**, **Curry, JA**, **Du, P**, **Girard, E**, **Haugen, J-E**, **Jacob, D**, **Jones, CG**, **Køltzow, M**, **Laprise, R**, **Lynch, AH**, **Pfeifer, S**, **Serreze, MC**, **Shaw, MJ**, **Tjernstrom, M**, **Wyser, K**, **Zagar, M.** 2006. Evaluation of an ensemble of Arctic regional climate models: Spatiotemporal fields during the SHEBA year. *Climate Dynamics* **26**: 459–472. DOI: http://dx.doi.org/10.1007/s00382-005-0095-3.

Salas Mélia, D. 2002. A global coupled sea ice–ocean model. *Ocean Modelling* **4**(2): 137–172. DOI: http://dx.doi.org/10.1016/S1463-5003(01)00015-4.

Sedlar, J, **Tjernström, M**, **Rinke, A**, **Orr, A**, **Cassano, J**, **Fettweis, X**, **Heinemann, G**, **Seefeldt, M**, **Solomon, A**, **Matthes, H**, **Phillips, T**, **Webster, S.** 2020. Confronting Arctic troposphere, clouds, and surface energy budget representations in regional climate models with observations. *Journal of Geophysical Research: Atmosphere* **125**: e2019JD031783. DOI: http://dx.doi.org/10.1029/2019jd031783.

Sengupta, M, **Andreas, A**, **Habte, A**, **Kutchenreiter, M**, **Reda, I**, **Xie, Y**, **Gotseff, P**. 2021. Ground Radiometers on Stand for Upwelling Radiation (GNDRAD60 S). Atmospheric Radiation Measurement (ARM) User Facility. DOI: http://dx.doi.org/10.5439/1025192. Accessed July 15, 2022.

Shupe, MD. 2022. ShupeTurner cloud microphysics product. ARM Mobile Facility (MOS) MOSAiC (Drifting Obs—Study of Arctic Climate). DOI: http://dx.doi.org/10.5439/1871015. Accessed July 15, 2022.

Shupe, MD, **Intrieri, JM.** 2004. Cloud radiative forcing of the Arctic surface: The influence of cloud properties,

surface albedo, and solar zenith angle. *Journal of Climate* **17**(3): 616–628. DOI: http://dx.doi.org/10.1175/1520-0442(2004)017<0616:crfota>2.0.CO;2.

Shupe, MD, **Rex, M**, **Blomquist, B**, **Persson, POG**, **Schmale, J**, **Uttal, T**, **Althausen, D**, **Angot, H**, **Archer, S**, **Bariteau, L**, **Beck, I**, **Bilberry, J**, **Bussi, S**, **Buck, C**, **Boyer, M**, **Brasseur, Z**, **Brooks, IM**, **Calmer, R**, **Cassano, J**, **Castro, V**, **Chu, D**, **Costa, D**, **Cox, CJ**, **Creamean, J**, **Crewell, S**, **Dahlke, S**, **Damm, E**, **de Boer, G**, **Deckelmann, H**, **Dethloff, K**, **Dütsch, M**, **Ebell, K**, **Ehrlich, A**, **Ellis, J**, **Engelmann, R**, **Fong, AA**, **Frey, MM**, **Gallagher, MR**, **Ganzeveld, L**, **Gradinger, R**, **Graeser, J**, **Greenamyer, V**, **Griesche, H**, **Griffiths, S**, **Hamilton, J**, **Heinemann, G**, **Helmig, D**, **Herber, A**, **Heuzé, C**, **Hofer, J**, **Houchens, T**, **Howard, D**, **Inoue, J**, **Jacobi, H-W**, **Jaiser, R**, **Jokinen, T**, **Jourdan, O**, **Jozef, G**, **King, W**, **Kirchgaessner, A**, **Klingebiel, M**, **Krassovski, M**, **Krumpen, T**, **Lampert, A**, **Landing, W**, **Laurila, T**, **Lawrence, D**, **Loose, B**, **Lonardi, M**, **Lüpkes, C**, **Maahn, M**, **Macke, A**, **Maslowski, W**, **Marsay, C**, **Maturilli, M**, **Mech, M**, **Morris, S**, **Moser, M**, **Nicolaus, M**, **Ortega, P**, **Osborn, J**, **Pätzold, F**, **Perovich, DK**, **Petäjä, T**, **Pilz, C**, **Pirazzini, R**, **Posman, K**, **Powers, H**, **Pratt, KA**, **Preußer, A**, **Quéléver, L**, **Radenz, M**, **Rabe, B**, **Rinke, A**, **Sachs, T**, **Schulz, A**, **Siebert, H**, **Silva, T**, **Solomon, A**, **Sommerfeld, A**, **Spreen, G**, **Stephens, M**, **Stohl, A**, **Svensson, G**, **Uin, J**, **Viegas, J**, **Voigt, C**, **von der Gathen, P**, **Wehner, B**, **Welker, JM**, **Wendisch, M**, **Werner, M**, **Xie, Z**, **Yue, F**. 2022. Overview of the MOSAiC expedition: Atmosphere. *Elementa: Science of the Anthropocene* **10**(1). DOI: http://dx.doi.org/10.1525/elementa.2021.00060.

Shupe, MD, **Turner, DD**, **Zwink, A**, **Thieman, MM**, **Mlawer, EJ**, **Shippert, T.** 2015. Deriving Arctic cloud microphysics at Barrow, Alaska: Algorithms, results, and radiative closure. *Journal of Applied Meteorology and Climatology* **54**: 1675–1689. DOI: http://dx.doi.org/10.1175/JAMC-D-15-0054.1.

Simjanovski, D, **Girard, E**, **Du, P**. 2011. An evaluation of Arctic cloud and radiation processes simulated by the limited-area version of the global multiscale environmental model (GEM-LAM). *Atmosphere-Ocean* **49**(3): 219–234. DOI: http://dx.doi.org/10.1080/07055900.2011.604266.

Solomon, A, **Intrieri, J**, **Cox, C**, **Persson, O**, **de Boer, G**, **Hughes, M**, **Capotondi, A**, **Shupe, M**. 2023. Evaluation of the NOAA Experimental Coupled Arctic Forecast System (CAFS). *The Cryosphere*, in preparation.

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

Tjernström, M, **Svensson, G**, **Magnusson, L**, **Brooks, IM**, **Prytherch, J**, **Vüllers, J**, **Young, G**. 2021. Central Arctic weather forecasting: Confronting the ECMWF IFS with observations from the Arctic Ocean 2018 expedition. *Quarterly Journal of the Royal Meteorological Society* **147**: 1278–1299. DOI: http://dx.doi.org/10.1002/qj.3971.

Tjernström, M, **Zagar, M**, **Svensson, G**, **Cassano, J**, **Pfeifer, S**, **Rinke, A**, **Wyser, K**, **Dethloff, K**, **Jones, C**, **Semmler, T**, **Shaw, M**. 2005. Modelling the Arctic boundary layer: An evaluation of six ARCMIP regional-scale models using data from the Sheba project. *Boundary-Layer Meteorology* **117**(2): 337–381. DOI: http://dx.doi.org/10.1007/s10546-004-7954-z.

Tolstykh, MA, **Fadeev, RY**, **Shashkin, VV**, **Goyman, GS**, **Zaripov, RB**, **Kiktev, DB**, **Makhnorylova, SV**, **Mizyak, VG**, **Rogutov, VS**. 2018. Multiscale Global Atmosphere Model SL-AV: The results of medium-range weather forecasts. *Russian Meteorology and Hydrology* **43**: 773–779. DOI: http://dx.doi.org/10.3103/S1068373918110080.

Uttal, T, **Curry, JA**, **McPhee, MG**, **Perovich, DK**, **Moritz, RE**, **Maslanik, JA**, **Guest, PS**, **Stern, HL**, **Moore, JA**, **Turenne, R**, **Heiberg, A**, **Serreze, MC**, **Wylie, DP**, **Persson, OG**, **Paulson, CA**, **Halle, C**, **Morison, JH**, **Wheeler, PA**, **Makshtas, A**, **Welch, H**, **Shupe, MD**, **Intrieri, JM**, **Stamnes, K**, **Lindsay, RW**, **Pinkel, R**, **Pegau, WS**, **Stanton, TP**, **Grenfeld, TC**. 2002. The surface heat budget of the Arctic. *Bulletin of the American Meteorological Society* **83**: 255–275. DOI: http://dx.doi.org/10.1175/1520-0477(2002)083<0255:SHBOTA>2.3.CO;2.

Wyser, K, **Jones, CG**, **Du, P**, **Girard, E**, **Willén, U**, **Cassano, J**, **Christensen, JH**, **Curry, JA**, **Dethloff, K**, **Haugen, J-E**, **Jacob, D**, **Køltzow, M**, **Laprise, R**, **Lynch, A**, **Pfeifer, S**, **Rinke, A**, **Serreze, M**, **Shaw, MJ**, **Tjernstroüm, M**, **Zagar, M**. 2008. An evaluation of Arctic cloud and radiation processes during the SHEBA year: Simulation results from eight Arctic regional climate models. *Climate Dynamics* **30**: 203–223. DOI: http://dx.doi.org/10.1007/s00382-007-0286-1.

Zängl, G, **Reinert, D**, **Ripodas, P**, **Baldauf, M**. 2015. The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society* **141**: 563–579. DOI: https://dx.doi.org/10.1002/qj.2378.

How to cite this article: Solomon, A, Shupe, MD, Svensson, G, Barton, NP, Batrak, Y, Bazile, E, Day, JJ, Doyle, JD, Frank, HP, Keeley, S, Remes, T, Tolstykh, M. 2023. The winter central Arctic surface energy budget: A model evaluation using observations from the MOSAiC campaign. *Elementa: Science of the Anthropocene* 11(1). DOI: https://doi.org/10.1525/elementa.2022.00104

Domain Editor-in-Chief: Detlev Helmig, Boulder AIR LLC, Boulder, CO, USA

Associate Editor: Joël Savarino, Laboratoire de Glaciologie et Géophysique de l'Environnement, CNRS/Grenoble University, Saint-Martin d'Hères, France

Knowledge Domain: Atmospheric Science

Part of an Elementa Special Feature: The Multidisciplinary Drifting Observatory for the Study of Arctic Climate (MOSAiC)

Published: April 19, 2023 **Accepted:** February 08, 2023 **Submitted:** August 10, 2022

ELEMENTA Science of the Anthropocene

Elem Sci Anth is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS