ELEMENTA
Science of the Anthropocene

RESEARCH ARTICLE

Evaluation of the Coupled Arctic Forecast System's representation of the Arctic atmospheric boundary layer vertical structure during MOSAiC

Gina Jozef[1,2,3,]*, John J. Cassano[1,2,3], Amy Solomon[2,4], Janet Intrieri[4], and Gijs de Boer[2,4,5]

Observations from the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) were used to evaluate the Coupled Arctic Forecast System (CAFS) model's ability to simulate the atmospheric boundary layer (ABL) structure in the central Arctic. MOSAiC observations of the lower atmosphere from radiosondes, downwelling longwave radiation (LWD) from a pyranometer, and near-surface wind conditions from a meteorological tower were compared to 6-hourly CAFS output. A self-organizing map (SOM) analysis reveals that CAFS reproduces the range of stability structures identified by the SOM trained with MOSAiC observations of virtual potential temperature (θ_v) profiles, but not necessarily with the correct frequency or at the correct time. Additionally, the wind speed profiles corresponding to a particular θ_v profile are not consistent between CAFS and the observations. When categorizing profiles by static stability, it was revealed that CAFS simulates all observed stability regimes, but overrepresents the frequency of near-surface strong stability, and underrepresents the frequency of strong stability between the top of the ABL and 1 km. The 10 m wind speeds corresponding to each stability regime consistently have larger values in CAFS versus observed, and this offset increases with decreasing stability. Whether LWD is over or underestimated in CAFS is dependent on stability regime. Both variables are most greatly overestimated in spring, leading to the largest near-surface θ_v bias, and the greatest underrepresentation of strong stability in spring. The results of this article serve to highlight the positive aspects of CAFS for representing the ABL and reveal impacts of misrepresentations of physical processes dictating energy, moisture, and momentum transfer in the lower troposphere on the simulation of central Arctic ABL structure and stability. This highlights potential areas for improvement in CAFS and other numerical weather prediction models. The SOM-based analysis especially provides a unique opportunity for process-based model evaluation.

1. Introduction

Accurate modeling of the central Arctic atmospheric boundary layer (ABL) is crucial for the accuracy of Arctic short-term weather forecasts as well as climate projections, each of which is important for varying reasons. The central Arctic is warming about 4 times faster than the rest of the planet (Rantanen et al., 2022), a phenomenon known as Arctic

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

[2] Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, CO, USA

[3] National Snow and Ice Data Center, University of Colorado Boulder, Boulder, CO, USA

[4] NOAA Physical Sciences Laboratory, Boulder, CO, USA

[5] Integrated Remote and In Situ Sensing, University of Colorado Boulder, Boulder, CO, USA

* Corresponding author:
Email: gina.jozef@colorado.edu

amplification (Serreze and Francis, 2006). One important repercussion of Arctic amplification is rapid sea ice loss (Serreze and Barry, 2011; Stroeve et al., 2012; Cohen et al., 2020; Screen, 2021), which creates the potential for increased human activity in the central Arctic, such as for scientific expeditions, personal exploration, or commercial shipping (Melia et al., 2016). For such activities, having accurate short-term weather forecasts from numerical weather prediction (NWP) models is important for maintaining safety during operations. Additionally, the central Arctic plays an important role in the global climate system through local thermodynamics (Tomas et al., 2016), and its contribution to the meridional temperature gradient which supports tropospheric westerly winds along the poleward flank of the jet stream (Deser et al., 2015; Francis and Vavrus, 2015; Labe et al., 2020). Thus, it is important that the central Arctic is properly simulated in climate models so that future change under Arctic amplification can best be predicted.

The ABL is the turbulent lowest part of the atmosphere directly interacting with earth's surface (Stull, 1988; Marsik et al., 1995), the structure of which dictates the transfer of energy, moisture, and momentum between the earth's surface and overlying atmosphere (Brooks et al., 2017). Properly simulating the Arctic ABL structure has historically been a challenge for NWP and climate models. As such, improvements are needed in the representation of near-surface ABL height and stability in NWP and climate models in order to produce accurate forecasts of short-term weather conditions or long-term climate conditions, respectively. In the current study, we evaluate the ability of the National Oceanic and Atmospheric Administration (NOAA) Coupled Arctic Forecast System (CAFS), a short-term NWP model (NOAA Physical Sciences Laboratory, n.d.), to accurately represent the Arctic ABL observed over the central Arctic ice pack between October 2019 and September 2020, during the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC; Shupe et al., 2020). CAFS was chosen for this study, as NOAA had produced CAFS simulations for the full MOSAiC year with the goal of using MOSAiC observations to improve CAFS, and through this work, that goal can be supported. While climate model evaluation is not included in this article, the results of this study may be extrapolated to inform potential areas of improvement for climate models.

Previous studies, including those analyzing MOSAiC data, have shown that the Arctic ABL over sea ice is often stably stratified, with height less than 200 m, and sometimes less than 50 m (Esau and Sorokina, 2010; Jozef et al., 2024). Stable stratification often occurs when there is a negative downwelling longwave radiation (LWD) balance at the surface (i.e., in the absence of clouds) and when winds are light (Stull, 1988; Banta, 2008). Stable stratification can also occur when warm air advection transports layers of warm air over the colder near-surface environment (particularly common in Arctic summer; Tjernström et al., 2005). With similar frequency, the central Arctic ABL can be well-mixed and capped with a strong potential temperature inversion (Persson et al., 2002; Tjernström and Graversen, 2009; Esau and Sorokina, 2010; Jozef et al., 2023a; Jozef et al., 2024). Various thermodynamic and mechanical processes impact turbulence production in the Arctic ABL, contributing to the development of the frequent well-mixed layers separating the sea ice from the elevated inversion (Brooks et al., 2017). Such thermodynamic processes include advection of cold air over a warmer surface (Vihma et al., 2005), buoyant turbulence generation through emission of surface energy fluxes from open water regions such as leads (Lüpkes et al., 2008), turbulent mixing below cloud base from cloud-top radiative cooling (Tjernström et al., 2004; Chechin et al., 2023), or enhanced LWD from low-level clouds (Wang et al., 2001; Huang et al., 2021; Li et al., 2023). Such mechanical processes include the interaction between the atmosphere and surface roughness features such as ice edges and ridges (Andreas et al., 2010) or oceanic waves (Jenkins et al., 2021) and wind shear below the core of a low-level jet (Banta et al., 2003; Brooks et al., 2017).

Accurate representation of strong stability in the ABL has historically been a challenge for NWP and climate models (Tjernström et al., 2005; Cuxart et al., 2006; Steeneveld et al., 2006; Holtslag et al., 2013; Nigro et al., 2017; Schneider et al., 2021), which is owed in part to inadequate vertical resolution to resolve the often shallow stable ABL (Tjernström et al., 2005; Banta et al., 2007; Nigro et al., 2017). However, shortcomings in the representation of near-surface Arctic atmospheric stability can also be attributed to misrepresentation of the physical processes dictating energy, moisture, and momentum fluxes in the ABL (Pithan et al., 2016). As the vertical resolution of models improves, it becomes more evident that biases can be attributed to improper representation of such physical processes, and so understanding what processes and relationships a model is struggling to accurately represent helps to inform the adaptation of current ABL and surface-layer parameterizations (Baklanov et al., 2011).

One factor impacting the simulation of near-surface stability in NWP and climate models is the surface radiation budget (Tjernström et al., 2005; Steeneveld et al., 2006; Nigro et al., 2017) which, in the polar regions, is driven by longwave radiation and is largely forced by atmospheric moisture content and cloud cover (Wang et al., 2001; Persson et al., 2002; Tjernström et al., 2004; Chechin et al., 2023). Excessive downwelling surface radiative fluxes lead a model to underrepresent near-surface stability (Persson et al., 2002). Another reason why NWP and climate models may underrepresent the strength of near-surface stability is because the ABL parameterization is generating too much mechanical mixing (Cuxart et al., 2006; Holtslag et al., 2013). Such enhanced mechanical mixing causes a decrease in near-surface stability and can be due to an overestimation of near-surface wind speeds (Tjernström et al., 2005; Schneider et al., 2021). Conversely, an underestimation of wind speeds can lead to an overrepresentation of the strength of near-surface stability (Cuxart et al., 2006).

In this study, CAFS is evaluated based on its ability to accurately represent Arctic ABL profiles of virtual potential temperature (θ_v), wind speed, and stability. Possible sources of errors in the simulated profiles will be assessed by comparing CAFS model output to observations of meteorological state and surface radiative fluxes from MOSAiC. The questions guiding this study are as follows: Is CAFS able to recreate the ABL vertical structures and stability regimes present in the MOSAiC observations? Does CAFS recreate these structures and stability regimes with the correct frequency, or at the correct time? How do the radiative and mechanical forcings for the various stability regimes in CAFS differ from those observed?

To compare the vertical structures present in CAFS to the observations, as well as their relative frequencies and timing of occurrence, we assign CAFS profiles to the self-organizing map (SOM) patterns presented in Jozef et al. (2024), which illustrate the range in ABL θ_v structure throughout the MOSAiC year. Next, we compare the frequency distribution of stability regimes from CAFS versus the observed distribution. Lastly, we compare properties influencing ABL stability (e.g., LWD and near-surface wind speed) as simulated in CAFS versus observed quantities in
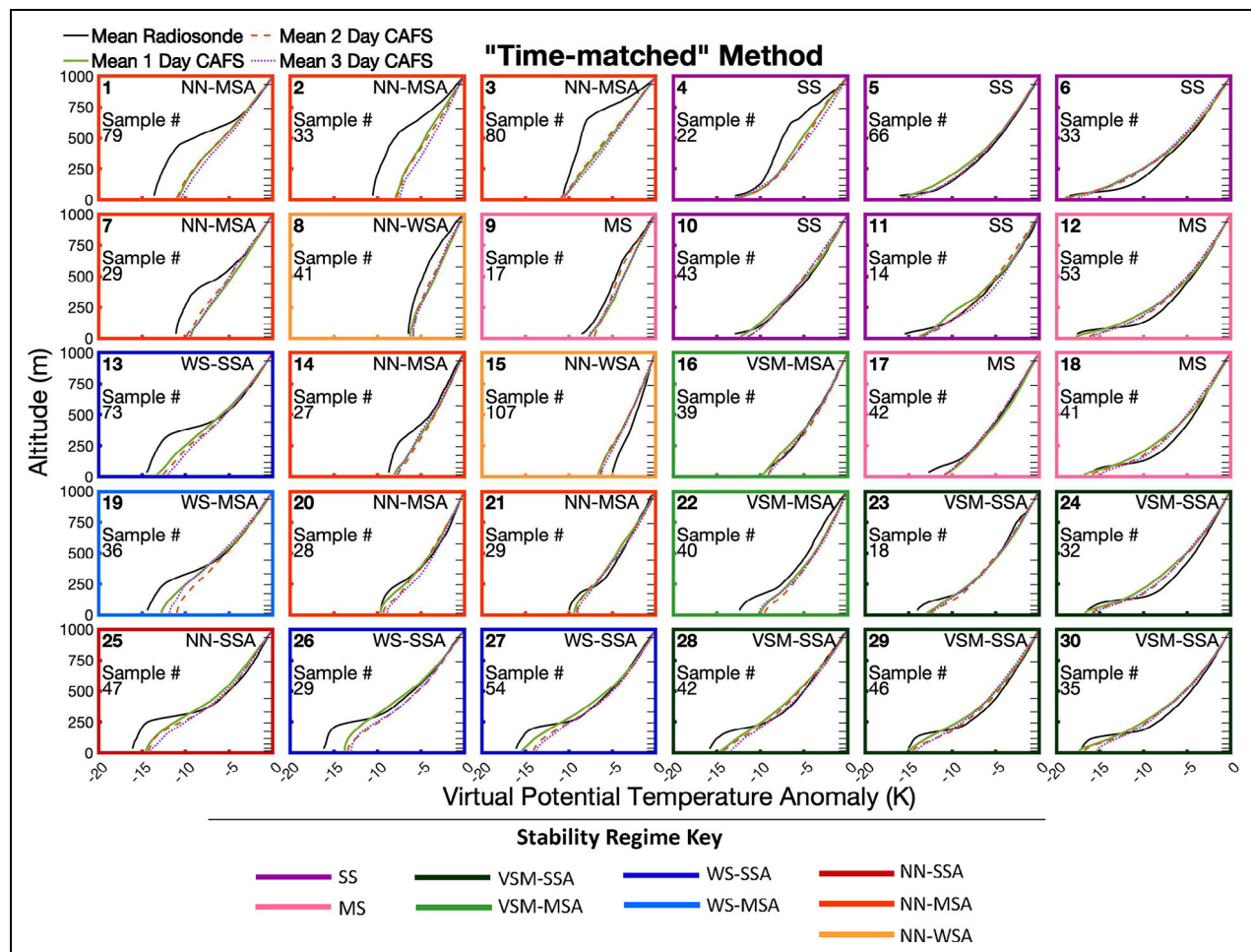
Figure 1. Radiosonde and Coupled Arctic Forecast System (CAFS) profiles assigned to each self-organizing map (SOM) pattern via the "time-matched" method. Profiles of mean virtual potential temperature (θ_v) anomaly for all radiosonde profiles (black solid) and all CAFS 1-day (green solid), 2-day (orange dashed), and 3-day (purple dotted) forecast profiles. The bold number in the upper left-hand corner of each subplot is the number of that SOM pattern (1 through 30), the number on the left side of each subplot is the number of radiosonde profiles (and thus the number of CAFS profiles) which map to that pattern, and the letters in the upper right-hand corner of each subplot indicates that pattern's stability regime (**Table 2**). Stability regime is also indicated by the color of the border for each subplot, following the colors given in the "Stability Regime Key." The horizontal tick marks along the right *y*-axis of each subplot indicate the CAFS native resolution.

2. Methods

2.1. Observational data from MOSAiC

Observational data used in this study were collected during MOSAiC, a year-long icebreaker-based expedition lasting from September 2019 through October 2020, during which the Research Vessel *Polarstern* (Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 2017) was frozen into the central Arctic Ocean sea ice and was set to drift passively across the central Arctic for the entire year (Jozef et al., 2024). To determine the vertical structure of the lower atmosphere, we use data from the balloon-borne Vaisala RS41 radiosondes, launched from the stern deck of the *Polarstern* (approximately 12 m above sea level). These

the context of season and stability regimes to determine where CAFS may be representing physical processes correctly or incorrectly and reveal potential areas for improvement.

launches occurred at least 4 times per day (every 6 h), typically at 05, 11, 17, and 23 UTC (Coordinated Universal Time) (Maturilli et al., 2021), however not all radiosonde observations were used for evaluation of CAFS (see Section 2.2). The locations of all MOSAiC radiosonde launches can be found in figure 1 of Jozef et al. (2023b), though methods were employed independent of radiosonde launch location. We use the level 2 Vaisala-processed radiosonde product (Maturilli et al., 2021) for this analysis, as the level 2 product is found to be more reliable in the lower troposphere than the GRUAN-processed level 3 product (Maturilli et al., 2022). See the abstracts for the level 2 (Maturilli et al., 2021) and level 3 (Maturilli et al., 2022) data for explanation of the difference between the 2 options (level 1 data are not publicly available so this was not considered). The radiosondes sample with a frequency of 1 Hz and ascend at a rate of approximately 5 m s^{-1}, resulting in a vertical resolution of about 5 m throughout the ascent. From the radiosondes, we

Table 1. Instrument brand and model name and uncertainty for each variable used in this study

Platform	Variable	Instrumentation	Uncertainty
Radiosonde	Pressure	Vaisala RS41-SGP	1.0 hPa (>100 hPa), 0.6 hPa (<100 hPa)
	Temperature		0.3°C (<16 km), 0.4°C (>16 km)
	Relative humidity		4%
	Wind speed		0.15 m s^{-1}
	Wind direction		2°
Radiation station	Downwelling longwave radiation (LWD)	Eppley Precision Infrared Radiometer	2.6 W m^{-2}
Met tower	10 m wind speed	Metek uSonic-Cage MP sonic anemometer	0.3 m s^{-1}

utilize measurements of temperature, pressure, relative humidity, wind speed, and wind direction.

Before radiosonde profiles were analyzed, measurements were corrected to account for the occasional passage of the balloon through the *Polarstern*'s smoke plume, and for the local "heat island" resulting from the presence of the *Polarstern*, which were identified by visual inspection of the temperature profile, as described in Jozef et al. (2023b). To allow for a consistent bottom height for our ABL analysis, we only consider profiles in which there are reliable data at 35 m AGL (above ground level) and above (Jozef et al., 2024), and all data at the altitudes below 35 m were removed. Data were determined to be reliable based on comparison of temperature from the radiosonde to that from a 10 m meteorological tower (hereafter "met tower"; Cox et al., 2023a) at the same time. Removing data below 35 m also helps in removing faulty wind measurements that occur as a result of flow distortion around the ship (Berry et al., 2001). Ultimately, 92% of radiosonde profiles were found to have credible measurements as low as 35 m, which removed 132 profiles from analysis. These 132 profiles are dispersed throughout the year, but many of them occurred over open ocean during active transit. As noted in Section 1, ABL height in the Arctic can be as shallow as less than 50 m, and thus by using data starting at 35 m, we accept that in some cases, we may miss a large portion of the ABL. However, when comparing radiosonde data above 35 m to near-surface measurements from the met tower, it was found that largely the key details in ABL structure just above the surface (e.g., inversion strength) were still detectable at 35 m.

For further analysis, radiosonde observations of temperature, relative humidity, wind speed components, and pressure were interpolated to a consistent vertical grid using 5 m spacing between 35 m and 1 km (linearly for temperature, relative humidity, and wind speed, and via the hypsometric equation for pressure). Then θ_v was calculated using the interpolated values of temperature, relative humidity, and pressure, and total wind speed was calculated using the interpolated values of zonal and meridional wind components. Additionally, 2 different versions of a virtual potential temperature gradient ($d\theta_v/dz$) profile were calculated for varying applications (see Sections 2.3.1 and 2.3.2).

ABL height from the radiosonde profiles was determined by identifying the first altitude in which the bulk Richardson number (Ri_b) exceeds a critical value of 0.5 and remains above the critical value for at least 20 consecutive meters, following the methods derived in Jozef et al. (2022) and applied in Jozef et al. (2023b). Ri_b was calculated using the following equation from Stull (1988):

$$Ri_b(z) = \frac{\left(\frac{g}{\overline{\theta_v}}\right)\Delta\theta_v\,\Delta z}{\Delta u^2 + \Delta v^2},$$

where g is acceleration due to gravity, $\overline{\theta_v}$ is the mean virtual potential temperature over the altitude range being considered, z is altitude, u is zonal wind speed, v is meridional wind speed, and Δ represents the difference over the altitude range used to calculate Ri_b throughout the profile. Ri_b profiles were created by calculating Ri_b across 30 m intervals in steps of 5 m (Jozef et al., 2023b). These methods identify the ABL height as the bottom of the elevated θ_v inversion (or the bottom of the layer of enhanced θ_v inversion strength) for moderately stable to near-neutral (NN) conditions and at the top of the most stable layer for conditions with a strong surface-based θ_v inversion (Jozef et al., 2023a).

Measurements of surface LWD come from an Eppley Precision Infrared Radiometer (Cox et al., 2023a) deployed as part of the radiation station on the sea ice near the *Polarstern* (Cox et al., 2023b), and measurements of 10 m wind speed come from the met tower situated adjacent to the radiometer (Cox et al., 2023b). Both aforementioned variables were determined corresponding to each radiosonde launch as the average over a period of ±5 min from each radiosonde launch (10 min total; Jozef et al., 2023b). **Table 1** lists the instrument name and manufacturer-provided uncertainty for each of the observational variables used in this study. For wind speed in particular, it is recognized that the true uncertainty is likely to be greater than that provided by the manufacturer. Hamilton et al. (2022) found a difference in wind speeds of less than 1 m s^{-1} based on the 95% confidence intervals of observations from both radiosondes and the DataHawk2 uncrewed aircraft system (which observed the atmosphere during MOSAiC between 5 m and 1 km). A summary of these and

additional atmospheric observations collected during MOSAiC can be found in Shupe et al. (2022).

2.2. Coupled Arctic Forecast System

CAFS (NOAA Physical Sciences Laboratory, n.d.) is an experimental coupled pan-Arctic (Arctic basin and surrounding coasts with ocean boundaries at the Fram and Bering Straits) short-term weather forecast model (Solomon et al., 2023) adapted from the Regional Arctic System Model (Maslowski et al., 2012). CAFS component models include the Weather Research and Forecasting (WRF3.6.1) atmospheric model, the Parallel Ocean Program (POP2) model, the Los Alamos Community Ice Model (CICE5.1), and the National Center for Atmospheric Research Community Land Model (CLM4.5) coupled using a regionalized version of the Community Earth System Model flux coupler (CPL7). CAFS is run at a 10 km horizontal resolution with 40 vertical levels, including 11 vertical levels between 12 m and 1 km (where vertical resolution below 1 km decreases from about 25 m to 200 m with increasing altitude, as shown in **Figure 1**). The model was initialized at 00 UTC each day from October 1, 2019, to September 12, 2020, with initial and lateral boundary conditions from the National Centers for Environmental Prediction Global Forecast System (GFS) weather forecast model (National Centers for Environmental Information, 2020), and run free for 10 days (though only forecasts out to 3 days were used in this analysis).

The longest spin-up time in a forecast model is for the cloud fields. To limit this impact, initialization of the cloud fields was done with fields from the first day of the previous day's forecast. Also, restarts for the ocean and sea ice are used as initial conditions. However, because of the different data sources used to initialize the coupled system, there is a discontinuity between the initial sea surface temperature, sea ice concentration, and the GFS analyses used to initialize the atmospheric model. It must additionally be noted that the current study was done using model output at the center of the domain so for forecasts of a few days the lateral boundary conditions have limited impact. Most atmospheric state fields are largely determined by model physics and dynamics within 6 h of initialization so forecasts out to 3 days are expected to largely reflect modeled features rather than initial conditions.

For this study, 6-hourly CAFS output was taken from the grid point closest to the location of the *Polarstern* throughout MOSAiC (Solomon and Jozef, 2024), as in Solomon et al. (2023). For comparison with the MOSAiC radiosonde profile observations (and single-level observations at the time of radiosonde launch), we identify the date and time in the CAFS output which is closest to the date and time of each radiosonde launch (difference between radiosonde observation and corresponding CAFS times are typically <1 h). For each radiosonde, we identify the corresponding CAFS data for a 1-day forecast (CAFS initialized 6–24 h before time of interest), a 2-day forecast (CAFS initialized 30–48 h before time of interest), and a 3-day forecast (CAFS initialized 54–72 h before time of interest) to allow for an assessment of changes in model performance as a function of forecast lead time. For example, for a radiosonde observation at 12 UTC on April 3, the CAFS 1-day forecast would be the 12th hour from the model run initialized at 00 UTC on April 3, the CAFS 2-day forecast would be the 36th hour from the model run initialized at 00 UTC on April 2, the CAFS 3-day forecast would be the 60th hour from the model run initialized at 00 UTC on April 1. Considering only radiosonde observations which have a 1-, 2-, and 3-day forecast given the temporal scope of the CAFS output, we retain 1,275 MOSAiC radiosonde profiles with corresponding CAFS 1-, 2-, and 3-day forecasts for analysis, which are evenly distributed throughout the year.

CAFS profile variables of interest to this study include temperature, specific humidity, pressure, and zonal and meridional wind speed. To best be compared with radiosonde observations, CAFS profiles were interpolated following the same vertical grid of 5 m spacing and same methods as were applied to the radiosonde data, between the lowest CAFS model level (approximately 12 m) and 1 km, such that the lowest interpolated altitude for the CAFS data was 15 m. For calculations and analysis, only interpolated CAFS data at 35 m and above were used. CAFS data down to 15 m are however still shown in figures depicting CAFS profiles to provide the context for how interpolated values above 15 m were determined. Interpolating to this grid ensures that all details of the radiosonde observations are retained but does not add any additional information to the lower resolution CAFS data. Profiles of θ_v, $d\theta_v/dz$ (see Sections 2.3.1 and 2.3.2) and wind speed from CAFS were calculated using the interpolated values. ABL height was calculated from the CAFS profiles, following the methods applied to the radiosonde observations, as discussed in Section 2.1. Relevant single-level variables from CAFS include surface LWD and 10 m zonal and meridional wind speed (from which total wind speed was calculated), to be compared to the radiation station and met tower observations.

2.3. Grouping observational and CAFS data
2.3.1. SOMs

Jozef et al. (2024) determined the range of lower atmospheric stability profiles which occurred in the central Arctic during MOSAiC using a SOM analysis. The SOM analysis leverages an artificial neural network approach that objectively identifies a user selected number of patterns in a training dataset (Kohonen, 2001; Cassano et al., 2006; Liu and Weisberg, 2011; Sheridan and Lee, 2011). This starts with a random initialization of patterns from the input data space, followed by an iterative comparison of the training data to these patterns, where the pattern to which the input data sample is most similar, and adjacent neighboring patterns, are modified to reduce the squared difference between the pattern and the input data sample. After thousands of training iterations, the result is a two-dimensional array of patterns called a SOM, in which similar patterns are grouped, and the squared difference between the training data and the patterns present in the SOM are minimized (Kohonen, 2001; Cassano et al., 2015; Dice and Cassano, 2022). The SOM algorithm tends to

arrange patterns such that more muted patterns are situated on the inside of the SOM and more extreme patterns are situated around the outside of the SOM. A SOM is an unsupervised neural network algorithm, but the user must specify the size of the resulting two-dimensional array (the number and orientation of patterns; Cassano et al., 2015) before training begins.

In Jozef et al. (2024), the SOM was trained with 1,377 $d\theta_v/dz$ profiles from the MOSAiC radiosonde observations (all profiles with good data at 35 m and above). When employing a SOM analysis, oftentimes the training dataset and applied dataset are one in the same, as was the case in Jozef et al. (2024). These profiles of $d\theta_v/dz$ in K $(100 \text{ m})^{-1}$ were calculated from differences between consecutive pairs of θ_v at the 5 m grid spacing, resulting in values of $d\theta_v/dz$ at 37.5 m, 42.5 m, 47.5 m, and so on, with the last value being at 997.5 m. CAFS profiles of $d\theta_v/dz$ for analysis with the SOM were calculated in this same way.

The SOM presented in Jozef et al. (2024) has 30 patterns with an orientation of 6×5 that span the full range of profile types in the training data. While other SOM sizes were tested, this number and orientation provided the best balance of having enough patterns to truly highlight the range in stability (which is missed with too few patterns) while not having several patterns that look nearly identical (which happens with too many patterns), and displays a smoothly varying, continuous depiction of the range of stability structures present in the data. Each radiosonde profile was "mapped" to the SOM (Cassano et al., 2016; Dice and Cassano, 2022) by finding the SOM pattern with the least squared difference when compared to the radiosonde profile. For further discussion on the methods for creating the MOSAiC radiosonde SOM, see Jozef et al. (2024).

2.3.2. Stability regime

To compare MOSAiC observations to CAFS output in the context of lower atmospheric static stability, we determine the stability regime of each radiosonde and CAFS profile using profiles of $d\theta_v/dz$ and ABL height following the methods of Jozef et al. (2024) and Jozef et al. (2023a). Stability regime identification was also applied to the SOM patterns based on the mean radiosonde profiles, as in Jozef et al. (2024), to provide the reader with a visual representation of vertical stability profiles associated with the different stability regimes and to demonstrate the variability in profiles within each stability regime.

For this application, $d\theta_v/dz$ was calculated across 30 m intervals in steps of 5 m between 35 m and 1 km and attributed to the center altitude of Δz (i.e., 35–65 m, 40–70 m, 45–75 m, and so on). An additional $d\theta_v/dz$ value calculated across a 15 m interval between 35 m and 50 m was included at the bottom of the profile, such that ultimately the $d\theta_v/dz$ profile has values at 42.5 m, 50 m, 55 m, 60 m AGL, and so on; $d\theta_v/dz$ was not calculated across 5 m intervals as was done for the SOM application, as that type of $d\theta_v/dz$ profile was found to be too noisy for the purposes of calculating representative stability regimes.

Per the methods of Jozef et al. (2024) and Jozef et al. (2023a), stability regime first depends on the $d\theta_v/dz$ profile value at 42.5 (the lowest altitude in the $d\theta_v/dz$ profile

created for this application), hereafter called the near-surface stability, where categories are strongly stable (SS), moderately stable (MS), weakly stable (WS), and NN. Next, for cases with a near-surface regime of WS or NN, the case is considered very shallow mixed (VSM) if the ABL height is less than 125 m (this separation was employed as there are different processes that would lead to a shallow vs. deep mixed layer). Lastly, stability aloft is determined for VSM, WS, and NN cases based on the maximum value in the $d\theta_v/dz$ profile above the ABL, but below 1 km, where the categories are strongly stable aloft (SSA), moderately stable aloft (MSA), and weakly stable aloft (WSA).

All resulting options for stability regime and the identifying thresholds are listed in **Table 2** and were determined based on the range of profiles seen in the SOM from Jozef et al. (2024) as well as from a similar SOM-based analysis of ABL profiles in Antarctica (Dice et al., 2023). This method development is discussed at length in Jozef et al. (2024). The color-coding in **Table 2** will be used to discern each regime henceforth, where regimes with the same near-surface stability are given the same color group, such that the lighter the shade is within that color group, the weaker the stability aloft is. The regimes are organized from strongest to weakest near-surface stability from top to bottom (where VSM is considered more stable than WS due to a shallower ABL). Within a given near-surface regime, the aloft regimes are also organized such that stability decreases from top to bottom.

Though the first altitude with data considered for determination of near-surface stability regime is 35 m, we are confident that the stability at this height is representative of the bulk ABL stability. Before settling on these methods, we tested using data down to 25 m for cases with good data at this height and found little difference in the stability regime classifications. We also compared the radiosonde profiles to the met tower data collected at the same time, and found the stability above 6 m to largely agree with that at 35 m. However, it is expected that in some very stable cases, the strength of stability below 35 m might exceed that which is observed above 35 m, but this largely would not ultimately affect the stability regime classification.

Both the SOM and stability regime methods were used to evaluate the CAFS model. The SOM provides a tangible demonstration of the range of stability structures in the CAFS data versus the radiosonde observations and reveals nuanced differences for both cases of different stability regimes, as well as cases of the same stability regime but different inversion height and strength (Sections 3.1.1 and 3.2.1). The stability regime method allows for a deeper analysis of the turbulent forcing mechanisms, which may explain differences between the simulations and the observations (Sections 3.1.2 and 3.2.2). Additionally, the stability regimes can be applied anywhere (e.g., Dice et al., 2023; Dice et al., 2024) while the SOM is only applicable to the site used to train the SOM.

2.3.3. "Time-matched" and "feature-matched" methods

The analyses used to initialize the CAFS model assimilate the MOSAiC soundings so are constrained at the initial

Table 2. Thresholds for near surface (first column) and aloft (third column) dθ_v/dz and ABL height (second column) used to identify stability regimes (fourth column) and their abbreviation (last column)[a]

dθ_v/dz at 42.5 m AGL	ABL Height	Max. dθ_v/dz Above ABL	Stability Regime	Abbreviation
$\geq$5 K (100 m)$^{-1}$	–	–	Strongly stable	SS
$\geq$1.75 K (100 m)$^{-1}$ <5 K (100 m)$^{-1}$	–	–	Moderately stable	MS
<1.75 K (100 m)$^{-1}$	<125 m	$\geq$5 K (100 m)$^{-1}$	Very shallow mixed—Strongly stable aloft	VSM-SSA
		$\geq$1.75 K (100 m)$^{-1}$ <5 K (100 m)$^{-1}$	Very shallow mixed—Moderately stable aloft	VSM-MSA
		<1.75 K (100 m)$^{-1}$	Very shallow mixed—Weakly stable aloft	VSM-WSA
$\geq$0.5 K (100 m)$^{-1}$ <1.75 K (100 m)$^{-1}$	$\geq$125 m	$\geq$5 K (100 m)$^{-1}$	Weakly stable—Strongly stable aloft	WS-SSA
		$\geq$1.75 K (100 m)$^{-1}$ <5 K (100 m)$^{-1}$	Weakly stable—Moderately stable aloft	WS-MSA
		<1.75 K (100 m)$^{-1}$	Weakly stable	WS
<0.5 K (100 m)$^{-1}$		$\geq$5 K (100 m)$^{-1}$	Near-neutral—Strongly stable aloft	NN-SSA
		$\geq$1.75 K (100 m)$^{-1}$ <5 K (100 m)$^{-1}$	Near-neutral—Moderately stable aloft	NN-MSA
		<1.75 K (100 m)$^{-1}$	Near-neutral—Weakly stable aloft	NN-WSA
		$\geq$0.5 K (100 m)$^{-1}$ <1.75 K (100 m)$^{-1}$	Near-neutral	NN

[a]This color-coding is used to discern each regime throughout the article. Regimes with the same near-surface stability are given the same color group, where a lighter shade indicates weaker stability aloft, and a darker shade indicates stronger stability aloft. The regimes are organized from strongest to weakest near-surface stability from top to bottom (where VSM is considered more stable than WS due to a shallower ABL). Within a given near-surface regime, the aloft regimes are also organized such that stability decreases from top to bottom.

time, but the model fields will diverge from the observations as a function of lead time. However for the relatively short-term (up to 3 day) forecasts assessed in this study, we expect that the model will successfully simulate the correct timing of weather features passing over the MOSAiC site. Thus, it is useful to understand both if CAFS generally simulates observed atmospheric features, and if it simulated these features at the correct time. Hence, the CAFS simulations are compared to the radiosonde observations through 2 methods: the "time-matched" method and the "feature-matched" method, each of which seek to answer different questions.

The "time-matched" method matches time-corresponding radiosonde and CAFS profiles to assess similarities and differences in ABL vertical structure, stability, and turbulent forcing mechanisms. This analysis answers the questions: Is CAFS able to simulate the correct ABL vertical structure and stability at the correct time? How do the radiative and mechanical forcings in CAFS differ from those observed for a given case? and How do errors in those forcings drive differences between the observed and simulated stability regimes? The "time-matched" method utilizes the SOM generated from the radiosonde profiles in Jozef et al. (2024) by assigning a CAFS profile to the SOM pattern to which its time-corresponding radiosonde observation is mapped (this may not be the pattern with the least squared difference when compared to the CAFS profile). The stability regime analysis is utilized in the "time-

matched" method by comparing stability regime and radiative and mechanical forcings from a given CAFS simulation to those from the time-corresponding radiosonde observation. This analysis allows us to evaluate CAFS's ability to simulate the observed features at the same time that they were observed.

The "feature-matched" method groups radiosonde and CAFS profiles separately based on their respective ABL vertical structure or stability to assess the similarities and differences in the turbulent forcing mechanisms that lead to given ABL structure and is therefore not concerned with whether the correct structures are simulated at the correct time. This analysis answers the questions: Is CAFS able to recreate ABL vertical structures and stability regimes present in the MOSAiC observations with the correct frequencies, regardless of whether the features are simulated at the correct time? How do the radiative and mechanical forcings in CAFS differ from those observed for a given stability regime? The "feature-matched" method utilizes the SOM by assigning a CAFS profile to the SOM pattern with the least squared difference when compared to the CAFS profile, above 35 m. The "feature-matched" method utilizes the stability regime analysis by assigning each CAFS profile to a stability regime based on the model-produced $d\theta_v/dz$ profile, while radiosonde observations are assigned to stability regimes based on their observed $d\theta_v/dz$ profile. This results in 2 sets of data for each stability regime, one for CAFS and the other for radiosondes, that do not necessarily have the same number of samples or samples at the same time. These two datasets are compared to assess differences in the frequency of stability regimes and in radiative and mechanical forcings for a given stability regime in CAFS versus in the observations.

For the SOM analysis, to determine how well the CAFS profiles match the observations for both the "time-matched" and "feature-matched" methods, bias and root-mean-squared difference (RMSD) were calculated by comparing the mean profiles from CAFS and the observations within each SOM pattern, across the entire SOM. These mean profiles are constructed from the mean value across all cases in a given SOM pattern at a given altitude. As such, if there are inversions occurring at different heights or with different strengths, the mean profile will smooth out these features. Conversely, if inversions are largely occurring at the same height with the same strength, the mean profile will highlight these details.

3. Results and discussion
3.1. "Time-matched" method
3.1.1. SOM analysis
The results of assigning the CAFS profiles (from the 1-, 2-, and 3-day forecasts) to the SOM patterns through the "time-matched" method are shown in **Figure 1** with the mean θ_v anomaly profiles (where the anomaly equals the θ_v at a given altitude minus that at 1 km) from the radiosonde and CAFS profiles in each SOM pattern. Supplementary Figure S1 shows the mean θ_v anomaly profiles from the radiosondes and CAFS 1-day forecast for each SOM pattern, including 25th to 75th percentiles (percentiles from the 2- and 3-day forecasts do not differ discernably

from 1-day forecast percentiles). It must be noted that the stability regime of a given SOM pattern was determined from the mean radiosonde profile for that pattern when considering all radiosonde observations from MOSAiC with good data at 35 m and above (see **Figure 2** from Jozef et al., 2024). Thus, the stability regime of a mean CAFS profile for a given SOM pattern may not match the stability regime of that SOM pattern, since the CAFS profiles were mapped to the SOM based on their timing, not based on their profile structure.

For the patterns with near-surface stability of NN, WS, or VSM (largely the left half and bottom 2 rows of the SOM), the mean CAFS profiles from all forecast periods show excessive stability in the near-surface layer (i.e., the layer is not as well-mixed as in the observations) and do not capture the strength of the elevated θ_v inversion present in the observations. Rather, these profiles exhibit a gradual increase in θ_v with altitude between the surface and 1 km. For the patterns with stability of SS or MS (the upper right-hand corner of the SOM), the mean CAFS profiles from all forecast periods do capture the surface-based θ_v inversion, but the strength of the inversions again is not as strong as in the observations. Additionally, for most of the SOM patterns, the CAFS profiles exhibit a warm bias in the lower levels when compared to θ_v at 1 km (i.e., θ_v in the lower levels is not as much as less than that at 1 km), compared to the observations, and thus weaker stability between the surface and 1 km than is observed. Overall, the results shown in **Figure 1** suggest that CAFS often does not simulate the details of the vertical stability structure of the observed θ_v profile types at the right time.

Reasons for the differences in profile shape between CAFS and the observations will be explored in Section 3.1.2. The 1-, 2-, and 3-day forecast periods perform similarly when compared to the radiosonde profiles, though the total bias and RMSD across the SOM when comparing the CAFS to the radiosonde profiles (see Supplementary Table S1) do increase with lead time (1-day forecast bias and RMSD are 0.4 K and 1.2 K, respectively, compared to 0.5 K and 1.3 K for the 2-day forecast and 0.6 K and 1.4 K for the 3-day forecast). This suggests that as lead time increases, CAFS evolves to states that differ more from reality. However, there is not always a consistent directional shift in bias across the different SOM patterns when increasing lead time (i.e., for most SOM patterns, the mean CAFS profile moves farther away from the mean radiosonde profile as lead time increases, but for some patterns, the mean CAFS profile moves closer with increasing lead time).

We additionally show the mean wind speed profiles from the radiosonde and CAFS profiles in each SOM pattern when assigning CAFS profiles to the SOM using the "time-matched" method in **Figure 2**. Supplementary Figure S2 shows the mean wind speed profiles from the radiosondes and CAFS 1-day forecast for each SOM pattern, including 25th to 75th percentiles (percentiles from the 2- and 3-day forecasts do not differ discernably from 1-day forecast percentiles). Across the SOM, the mean wind speed profiles from CAFS are generally very similar in shape and magnitude to those from the observations. Any
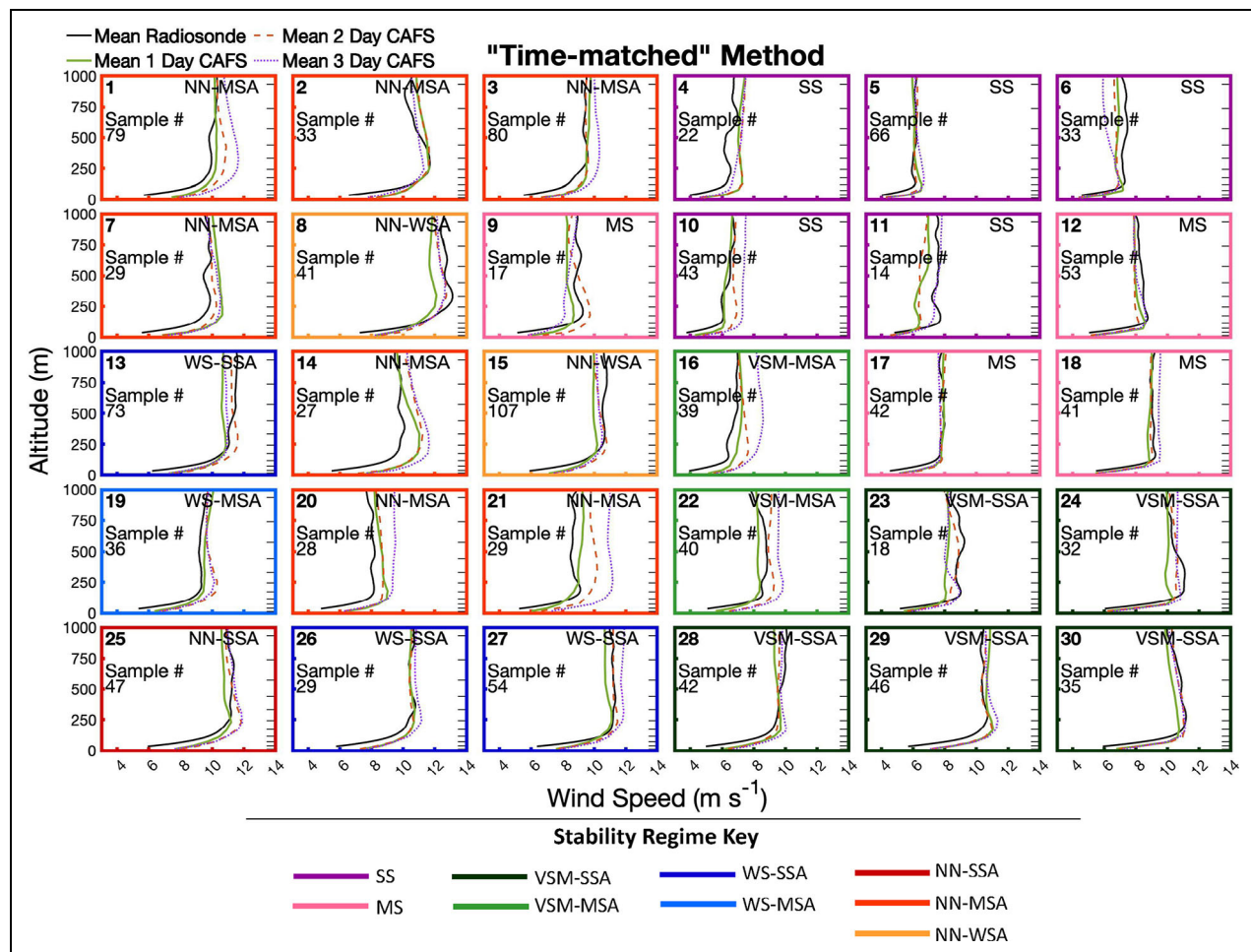
Figure 2. Radiosonde and Coupled Arctic Forecast System (CAFS) profiles assigned to each self-organizing map (SOM) pattern via the "time-matched" method. As **Figure 1** but for wind speed.

slight discrepancies more often manifest in overestimation than underestimation of wind speed from CAFS, especially at lower levels and near the altitude of the wind speed maximum. Referring to Supplementary Figure S2, there are no patterns in which the middle 50% of CAFS profiles do not overlap with the mean observation throughout the entire profile, aside from in the very lowest levels for some of the patterns (in which case the wind speeds from CAFS are too fast), indicating a high amount of agreement between the CAFS and radiosonde winds.

The above discussion leads to the conclusion that CAFS generally produces the right wind speed profiles at the right time, which suggests that CAFS simulates the correct large-scale pressure gradients at the right times to produce the proper wind speeds. This good performance cannot be simply attributed to the boundary conditions, as the CAFS grid point at the *Polarstern* is far from the model domain boundary, and forecast models have been found to simulate large-scale pressure gradients well on these timescales. However, since CAFS does not always create the right θ_v structure at the right time (**Figure 1**), this means that either the right wind speed profiles are driving erroneous θ_v structures, simulated θ_v structure is incorrect because of a different forcing mechanism (e.g., the height and characteristics of clouds), or there are other errors in

the model representation of the ABL, and associated processes (i.e., errors in θ_v structure are independent from turbulent forcing mechanisms). Additionally, while there is inconsistency across the SOM pattern as to which forecast period produces the wind speed profile most like the observations, total bias and RMSD (see Supplementary Table S1) do increase with lead time when comparing CAFS to radiosonde profiles (1-day forecast bias and RMSD are 0.0 m s^{-1} and 0.6 m s^{-1}, respectively, compared to 0.2 m s^{-1} and 0.7 m s^{-1} for the 2-day forecast and 0.2 m s^{-1} and 0.9 m s^{-1} for the 3-day forecast), suggesting that the 1-day forecast generally produces the most accurate wind speed profiles at the right time.

3.1.2. Stability regime analysis

In this section, we delve into why CAFS might simulate the right or wrong θ_v structure at a given time by comparing radiative and mechanical forcings in CAFS to the observations in the context of the stability regimes defined in Section 2.3.2. First we ask, when CAFS simulates the wrong stability regime, what stability regime does it tend to simulate instead? This question is explored in **Figure 3** which shows that when the radiosondes observed a certain stability regime (rows), with what frequency the CAFS 1-day forecast simulated each possible
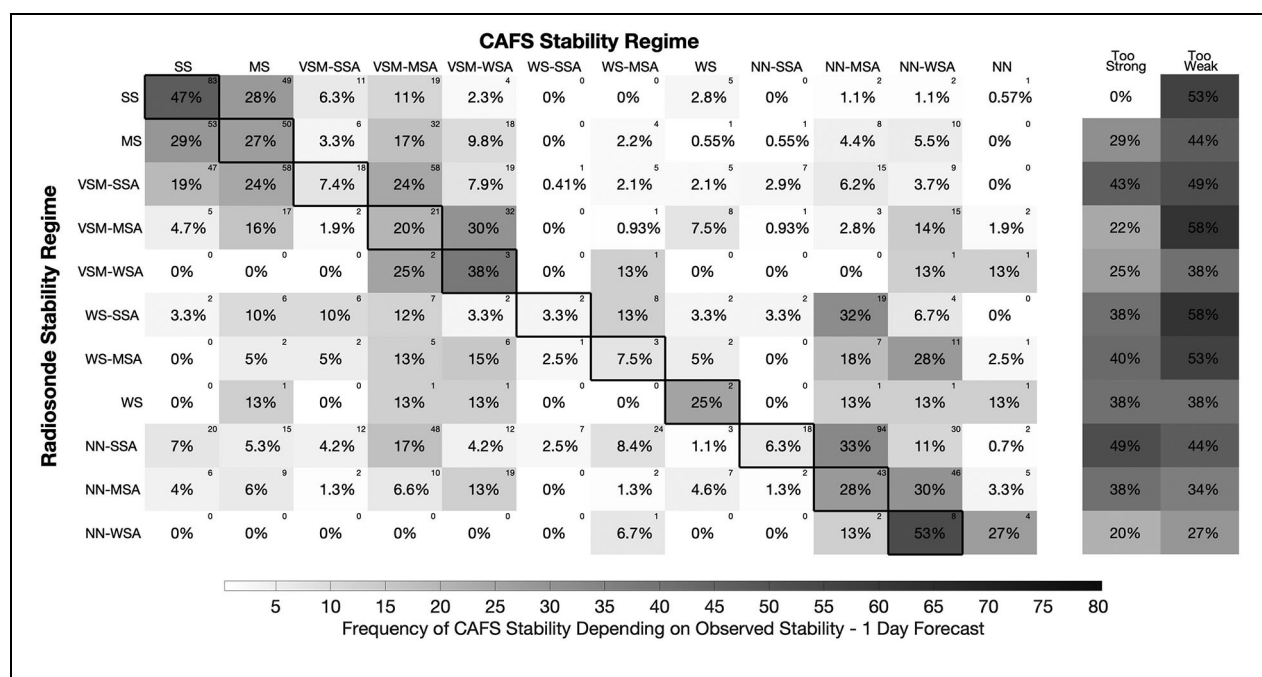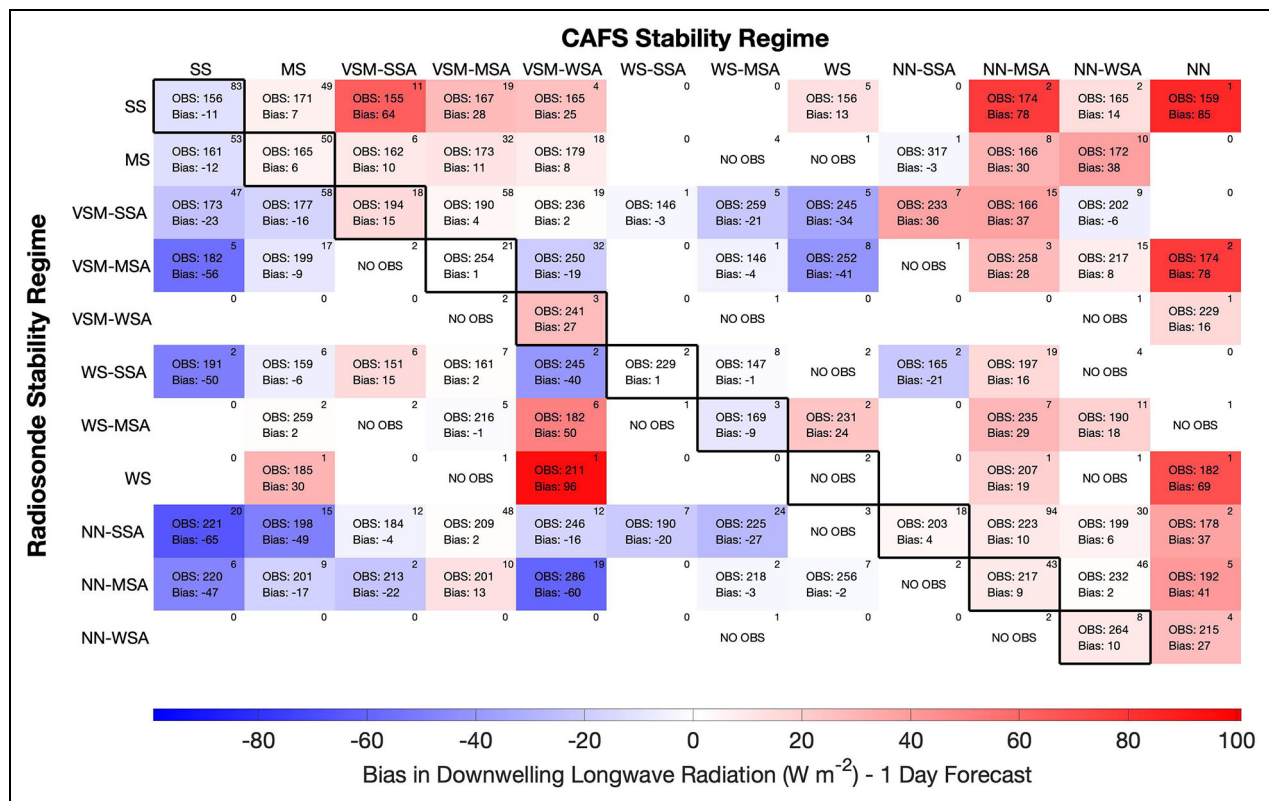
Figure 3. Coupled Arctic Forecast System (CAFS) versus radiosonde stability regimes. (Left) Grid plot showing, for each stability regime (**Table 2**) observed by the radiosondes (rows), what was the frequency of each stability regime simulated by the CAFS 1-day forecast (columns) at the corresponding time stamp. The grayscale color bar corresponds to the percent of CAFS cases in each stability regime for a given observed regime (also written in the center of each tile), where darker gray signifies a higher percent of cases, and each row adds up to 100%. Numbers in the upper right-hand corner indicate the number of cases in that tile. Black borders around the diagonal tiles indicate when CAFS simulated the same stability regime as was observed. (Right) Frequency with which CAFS simulates too strong of stability (sum of values to the left of the diagonal) or too weak of stability (sum of values to the right of the diagonal) for a given observed regime, following the same grayscale color bar.

stability regime at the same time (columns). For example, CAFS simulates MS for 28% of cases with the same time-stamps as when the radiosondes observed SS stability (row 1, column 2). The diagonal (boxes outlined in black) indicates when CAFS simulates the same stability regime as was observed. Boxes to the left of the diagonal indicate that CAFS stability is stronger than observed and boxes to the right of the diagonal indicate that CAFS stability is weaker than observed (these percentages are summed and shown the rightmost columns labeled "Too Strong" and "Too Weak" in **Figure 3**). No row for the radiosonde stability of NN is included, as there were no such observations. Supplementary Figure S3 shows the results for the CAFS 2- and 3-day forecasts.

For the 1-day forecast, CAFS is most consistent in simulating the correct stability regime when SS or NN-WSA was observed (47% and 53% of the time, respectively). These are the strongest (SS) and weakest (NN-WSA) stability regimes observed. This suggests that when conditions forcing atmospheric stability (e.g., surface radiation budget and wind) are on opposite sides of the spectrum, CAFS does a good job of simulating those endpoints, resulting in the correct simulation of associated stability regimes. However for stability regimes spanning the entire spectrum between SS and NN-WSA, CAFS is shown to struggle more frequently. While Jozef et al. (2023a) showed SS to be the most persistent regime, they found NN-WSA to be

one of the least persistent regimes so this phenomenon cannot be explained merely by the lifetime of the stability regime. For MS, CAFS most often simulates SS instead (29% of the time). For VSM-MSA, WS-SSA, WS-MSA, NN-SSA, and NN-MSA, the regime which is most often simulated instead of the correct regime is one with weaker stability. The regimes which most frequently simulate a stability that is several tiles away from the observed stability (i.e., CAFS stability is off by more than one regime) are WS-SSA and WS-MSA (in both cases, the simulated near-surface stability is NN and the stability aloft is one regime weaker).

For all observed regimes except WS, NN-SSA, and NN-MSA, the regime simulated by CAFS is generally less stable than observed (for WS, CAFS simulates too weak and too strong stabilities with equal frequencies). In most cases where the CAFS stability is too weak, it is stability aloft that is too weak, while near-surface stability is correct, which is consistent with the common warm bias in CAFS. The only exception is when near-surface stability is observed to be WS, in which case CAFS tends to simulate a near-surface stability of NN (except in a purely WS case). There are some regimes in the 2- and 3-day forecasts that are simulated at the right time more frequently in CAFS versus in the 1-day forecast (e.g., VSM-SSA, WS-SSA, and WS-MSA). However, on a broad scale when comparing the 2- and 3-day forecasts to the 1-day forecast, CAFS largely

CAFS Stability Regime

Radiosonde \ CAFS	SS	MS	VSM-SSA	VSM-MSA	VSM-WSA	WS-SSA	WS-MSA	WS	NN-SSA	NN-MSA	NN-WSA	NN
SS	OBS: 156, Bias: -11 (83)	OBS: 171, Bias: 7 (49)	OBS: 155, Bias: 64 (11)	OBS: 167, Bias: 28 (19)	OBS: 165, Bias: 25 (4)	(0)		OBS: 156, Bias: 13 (5)	(0)	OBS: 174, Bias: 78 (2)	OBS: 165, Bias: 14 (2)	OBS: 159, Bias: 85 (1)
MS	OBS: 161, Bias: -12 (53)	OBS: 165, Bias: 6 (50)	OBS: 162, Bias: 10 (6)	OBS: 173, Bias: 11 (32)	OBS: 179, Bias: 8 (18)	(0)	NO OBS (4)	NO OBS (1)	OBS: 317, Bias: -3 (1)	OBS: 166, Bias: 30 (8)	OBS: 172, Bias: 38 (10)	(0)
VSM-SSA	OBS: 173, Bias: -23 (47)	OBS: 177, Bias: -16 (58)	OBS: 194, Bias: 15 (18)	OBS: 190, Bias: 4 (58)	OBS: 236, Bias: 2 (19)	OBS: 146, Bias: -3 (1)	OBS: 259, Bias: -21 (5)	OBS: 245, Bias: -34 (5)	OBS: 233, Bias: 36 (7)	OBS: 166, Bias: 37 (15)	OBS: 202, Bias: -6 (9)	(0)
VSM-MSA	OBS: 182, Bias: -56 (5)	OBS: 199, Bias: -9 (17)	NO OBS (2)	OBS: 254, Bias: 1 (21)	OBS: 250, Bias: -19 (32)	(0)	OBS: 146, Bias: -4 (1)	OBS: 252, Bias: -41 (8)	NO OBS (1)	OBS: 258, Bias: 28 (3)	OBS: 217, Bias: 8 (15)	OBS: 174, Bias: 78 (1)
VSM-WSA	(0)	(0)	(0)	NO OBS (2)	OBS: 241, Bias: 27 (3)	(0)	NO OBS	(0)	(0)	(0)	NO OBS (1)	OBS: 229, Bias: 16 (1)
WS-SSA	OBS: 191, Bias: -50 (2)	OBS: 159, Bias: -6 (6)	OBS: 151, Bias: 15 (6)	OBS: 161, Bias: 2	OBS: 245, Bias: -40 (2)	OBS: 229, Bias: 1 (2)	OBS: 147, Bias: -1 (8)	NO OBS (2)	OBS: 165, Bias: -21 (2)	OBS: 197, Bias: 16 (19)	NO OBS (4)	(0)
WS-MSA	(0)	OBS: 259, Bias: 2 (2)	NO OBS (2)	OBS: 216, Bias: -1 (5)	OBS: 182, Bias: 50 (6)	NO OBS (1)	OBS: 169, Bias: -9 (3)	OBS: 231, Bias: 24	(0)	OBS: 235, Bias: 29 (7)	OBS: 190, Bias: 18 (11)	NO OBS (1)
WS	(0)	OBS: 185, Bias: 30 (1)	(0)	NO OBS	OBS: 211, Bias: 96 (1)	(0)	(0)	NO OBS (2)	(0)	OBS: 207, Bias: 19	NO OBS	OBS: 182, Bias: 69 (1)
NN-SSA	OBS: 221, Bias: -65 (20)	OBS: 198, Bias: -49 (15)	OBS: 184, Bias: -4 (12)	OBS: 209, Bias: 2 (48)	OBS: 246, Bias: -16 (12)	OBS: 190, Bias: -20 (7)	OBS: 225, Bias: -27 (24)	NO OBS (3)	OBS: 203, Bias: 4 (18)	OBS: 223, Bias: 10 (94)	OBS: 199, Bias: 6 (30)	OBS: 178, Bias: 37 (2)
NN-MSA	OBS: 220, Bias: -47 (6)	OBS: 201, Bias: -17 (9)	OBS: 213, Bias: -22 (2)	OBS: 201, Bias: 13 (10)	OBS: 286, Bias: -60 (19)	(0)	OBS: 218, Bias: -3 (2)	OBS: 256, Bias: -2 (7)	NO OBS	OBS: 217, Bias: 9 (43)	OBS: 232, Bias: 2 (46)	OBS: 192, Bias: 41 (5)
NN-WSA	(0)	(0)	(0)	(0)	(0)	(0)	NO OBS (1)	(0)	(0)	NO OBS	OBS: 264, Bias: 10 (2)	OBS: 215, Bias: 27 (8)

Bias in Downwelling Longwave Radiation (W m^{-2}) - 1 Day Forecast

Figure 4. Coupled Arctic Forecast System (CAFS) versus observed downwelling longwave radiation (LWD). Grid plot showing, for each stability regime (**Table 2**) observed by the radiosondes (rows), what was the bias (CAFS minus observational value) in LWD when the CAFS 1-day forecast simulated each stability regime (columns) at the corresponding time stamp. The color bar corresponds to the bias in CAFS. In the center of each tile, the observational mean (OBS) and bias are written. When there are cases which fall in a certain tile, but the LWD observations are missing for those cases, the text within that tile reads "NO OBS." When there are no cases in a certain tile, there is no text in the center of that tile. Numbers in the upper right-hand corner indicate the number of cases in that tile. Black borders around the diagonal tiles indicate when CAFS simulated the same stability regime as was observed.
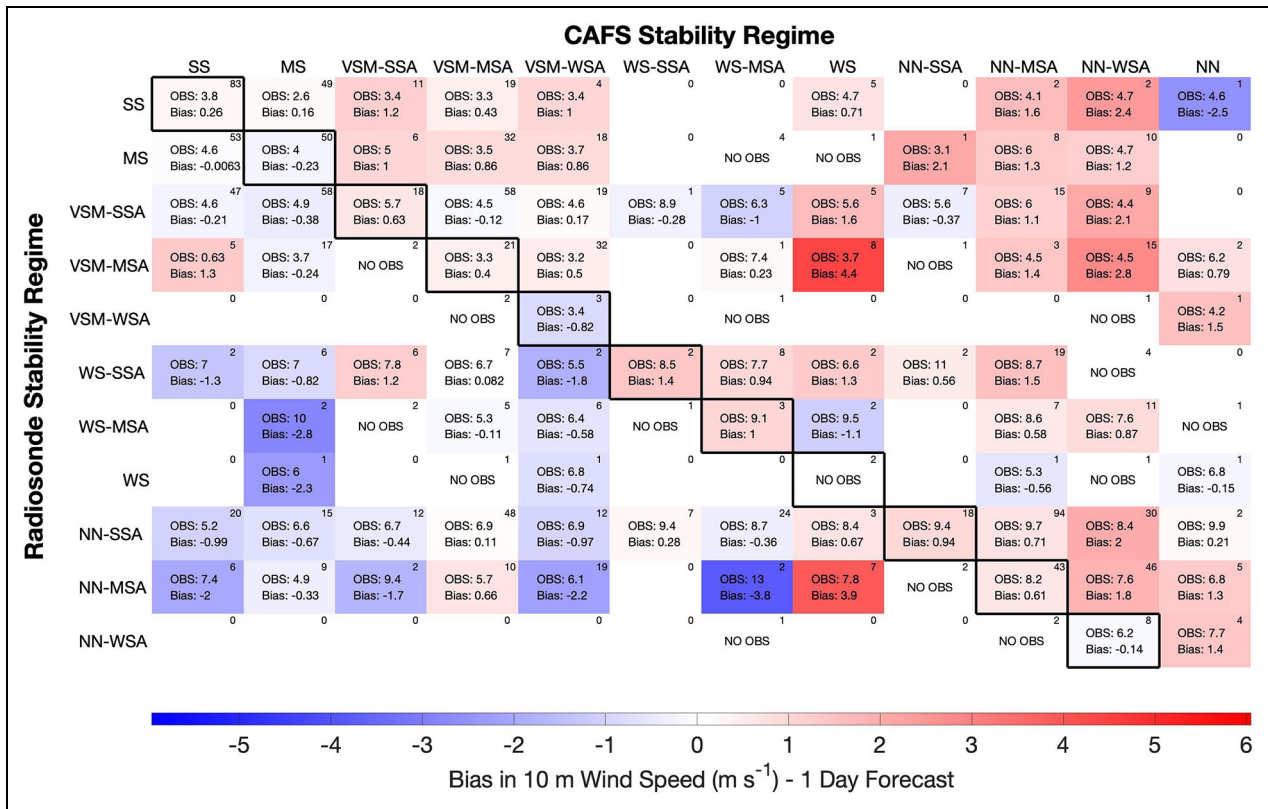
simulates the same regime as observed with lower frequencies and simulates too weak of stability with higher frequencies in the 2- and 3-day forecasts. Thus, we conclude that the CAFS stability regimes are more likely to be accurate with a 1-day forecast.

LWD is the atmosphere's contribution to the surface radiation budget and thus greatly impacts ABL structure and stability. Therefore, it is important to understand how LWD is represented in CAFS, and how the accuracy of stability regime simulation in CAFS varies with LWD. To demonstrate this, **Figure 4** shows the bias in LWD between CAFS and that observed at the time of radiosonde launch for each pair of stability regimes occurring at corresponding times (from **Figure 3**). Blue shading indicates that CAFS LWD was less than observed, and red shading indicates that CAFS LWD was greater than observed. When CAFS stability is weaker than observed, there is usually a positive bias in LWD, and when CAFS stability is stronger than observed, there is usually a negative bias in LWD. However, there are some exceptions, particularly when WS was observed and CAFS simulated VSM near-surface stability (in which the bias is often positive), and vice versa. Jozef et al. (2023a) found that VSM regimes occur under higher LWD conditions than WS regimes, and this result shows that CAFS properly simulates that relationship, albeit sometimes at the wrong time.

Generally, the greater the difference between the CAFS stability regime and the observed regime, the greater the bias in LWD (in both the positive and negative directions). Next, we will report the greatest positive and negative biases, considering boxes with a sample size of at least 5 (results with fewer than 5 samples are less likely to be significant). The greatest positive bias occurs when SS was observed and CAFS simulates VSM-SSA (bias of 64 W m^{-2}), highlighting the strong contribution of greater LWD for creating a VSM layer, found in Jozef et al. (2023a). The greatest negative bias occurs when NN-SSA was observed and CAFS simulates SS (bias of -65 W m^{-2}). Since these are toward the opposite sides of the stability regime spectrum, a large bias would be expected, and suggests that at these times CAFS may be simulating the presence of clouds while clouds were not observed. When CAFS simulates the correct regime, there is largely a positive bias for MS and the VSM and NN regimes, while SS and WS largely have a negative bias. Thus, CAFS does generally represent the correct relationships between stability and LWD, but

CAFS Stability Regime

Radiosonde \ CAFS	SS	MS	VSM-SSA	VSM-MSA	VSM-WSA	WS-SSA	WS-MSA	WS	NN-SSA	NN-MSA	NN-WSA	NN
SS	83 OBS: 3.8 Bias: 0.26	49 OBS: 2.6 Bias: 0.16	11 OBS: 3.4 Bias: 1.2	19 OBS: 3.3 Bias: 0.43	4 OBS: 3.4 Bias: 1	0		5 OBS: 4.7 Bias: 0.71	0	8 OBS: 4.1 Bias: 1.6	2 OBS: 4.7 Bias: 2.4	1 OBS: 4.6 Bias: -2.5
MS	53 OBS: 4.6 Bias: -0.0063	50 OBS: 4 Bias: -0.23	6 OBS: 5 Bias: 1	32 OBS: 3.5 Bias: 0.86	18 OBS: 3.7 Bias: 0.86	0	4 NO OBS	1 NO OBS	1 OBS: 3.1 Bias: 2.1	8 OBS: 6 Bias: 1.3	10 OBS: 4.7 Bias: 1.2	0
VSM-SSA	47 OBS: 4.6 Bias: -0.21	58 OBS: 4.9 Bias: -0.38	18 OBS: 5.7 Bias: 0.63	58 OBS: 4.5 Bias: -0.12	19 OBS: 4.6 Bias: 0.17	OBS: 8.9 Bias: -0.28	5 OBS: 6.3 Bias: -1	5 OBS: 5.6 Bias: 1.6	7 OBS: 5.6 Bias: -0.37	15 OBS: 6 Bias: 1.1	9 OBS: 4.4 Bias: 2.1	0
VSM-MSA	5 OBS: 0.63 Bias: 1.3	17 OBS: 3.7 Bias: -0.24	2 NO OBS	21 OBS: 3.3 Bias: 0.4	32 OBS: 3.2 Bias: 0.5	0	1 OBS: 7.4 Bias: 0.23	8 OBS: 3.7 Bias: 4.4	1 NO OBS	3 OBS: 4.5 Bias: 1.4	15 OBS: 4.5 Bias: 2.8	2 OBS: 6.2 Bias: 0.79
VSM-WSA	0	0	0	2 NO OBS	3 OBS: 3.4 Bias: -0.82	0	NO OBS	0	0	0	NO OBS	1 OBS: 4.2 Bias: 1.5
WS-SSA	2 OBS: 7 Bias: -1.3	6 OBS: 7 Bias: -0.82	6 OBS: 7.8 Bias: 1.2	7 OBS: 6.7 Bias: 0.082	5 OBS: 5.5 Bias: -1.8	2 OBS: 8.5 Bias: 1.4	8 OBS: 7.7 Bias: 0.94	2 OBS: 6.6 Bias: 1.3	2 OBS: 11 Bias: 0.56	19 OBS: 8.7 Bias: 1.5	4 NO OBS	0
WS-MSA	0	2 OBS: 10 Bias: -2.8	2 NO OBS	5 OBS: 5.3 Bias: -0.11	6 OBS: 6.4 Bias: -0.58	1 NO OBS	3 OBS: 9.1 Bias: 1	OBS: 9.5 Bias: -1.1	0	7 OBS: 8.6 Bias: 0.58	11 OBS: 7.6 Bias: 0.87	1 NO OBS
WS	0	1 OBS: 6 Bias: -2.3	0	1 NO OBS	1 OBS: 6.8 Bias: -0.74	0	0	NO OBS	0	1 OBS: 5.3 Bias: -0.56	NO OBS	1 OBS: 6.8 Bias: -0.15
NN-SSA	20 OBS: 5.2 Bias: -0.99	15 OBS: 6.6 Bias: -0.67	12 OBS: 6.7 Bias: -0.44	48 OBS: 6.9 Bias: 0.11	12 OBS: 6.9 Bias: -0.97	7 OBS: 9.4 Bias: 0.28	24 OBS: 8.7 Bias: -0.36	3 OBS: 8.4 Bias: 0.67	18 OBS: 9.4 Bias: 0.94	94 OBS: 9.7 Bias: 0.71	30 OBS: 8.4 Bias: 2	OBS: 9.9 Bias: 0.21
NN-MSA	6 OBS: 7.4 Bias: -2	9 OBS: 4.9 Bias: -0.33	2 OBS: 9.4 Bias: -1.7	10 OBS: 5.7 Bias: 0.66	19 OBS: 6.1 Bias: -2.2	0	2 OBS: 13 Bias: -3.8	7 OBS: 7.8 Bias: 3.9	2 NO OBS	43 OBS: 8.2 Bias: 0.61	46 OBS: 7.6 Bias: 1.8	5 OBS: 6.8 Bias: 1.3
NN-WSA	0	0	0	0	0	0	1 NO OBS	0	0	2 NO OBS	8 OBS: 6.2 Bias: -0.14	4 OBS: 7.7 Bias: 1.4

Bias in 10 m Wind Speed (m s^{-1}) - 1 Day Forecast

Figure 5. Coupled Arctic Forecast System (CAFS) versus observed 10 m wind speed. As **Figure 4** but for 10 m wind speed.

the magnitude of LWD which creates a certain stability regime is not always right, largely being too high for MS, VSM, and NN and too low for SS and WS.

For the 2- and 3-day forecasts (Supplementary Figure S4), the same overall results hold (i.e., that the bias is usually positive [negative] when CAFS stability is weaker [stronger] than observed, with largely the same direction of bias when CAFS simulates the correct regime). However, in the 2- and 3-day forecasts, there are more positive biases across the grid when CAFS undersimulates stability (mean bias of 17.8 W m^{-2} in the 1-day forecast vs. 28.4 and 23.5 W m^{-2} in the 2- and 3-day forecasts, respectively). This finding suggests that increasing lead time from 1 to 2 or 3 days leads to a greater overestimation of LWD.

Wind speed at 10 m AGL is a proxy for mechanical mixing, which impacts ABL stability. Therefore, it is important to understand the relationship between near-surface wind speed and stability in CAFS. Thus, we evaluate differences in the 10 m wind speed simulated by CAFS relative to observations for each combination of observed versus simulated stability regime, in a time-matched manner. This is depicted in **Figure 5** which is laid out like **Figure 4**, but color coded by the bias in 10 m wind speed between CAFS and observed at the time of radiosonde launch. As with LWD, when CAFS-simulated stability is weaker than that observed, there is usually a positive bias in 10 m wind speed, and when CAFS-simulated stability is stronger than that observed, there is usually a negative bias in 10 m wind speed. The exceptions are placed seemingly randomly around the grid, such that a consistent or explainable pattern is not found.

As for LWD, larger differences between CAFS and observed stability regimes are associated with greater wind speed biases (in both the positive and negative directions). Considering boxes with a sample size of at least 5, the greatest positive bias occurs when VSM-MSA was observed and CAFS simulates WS (bias of 4.4 m s^{-1}). The greatest negative bias occurs when NN-MSA was observed and CAFS simulates VSM-WSA (bias of -2.2 m s^{-1}). When CAFS simulates the correct regime, there is usually a small positive bias (average bias of 0.4 m s^{-1} considering boxes with sample size of at least 5). These results suggest that CAFS does generally represent the correct relationships between stability and near-surface wind speed, but the magnitude of 10 m wind speed which creates a certain stability regime is generally too high. As for LWD, the same overall results hold for the 2- and 3-day forecasts (Supplementary Figure S5), but there are more positive biases across the grid when CAFS simulates too weak of stability (mean bias of 0.9 m s^{-1} in the 1-day forecast vs. 1.5 and 1.3 m s^{-1} in the 2- and 3-day forecasts, respectively). This finding suggests that increasing lead time from 1 to 2 or 3 days leads to a greater overestimation of 10 m wind speed.

3.2. "Feature-matched" method
3.2.1. SOM analysis
The results of assigning CAFS profiles to SOM patterns through the "feature-matched" method are shown in **Figure 6** with mean θ_v anomaly profiles from radiosondes and CAFS for each SOM pattern. Supplementary Figure S6 shows mean θ_v anomaly profiles from radiosondes and CAFS 1-day
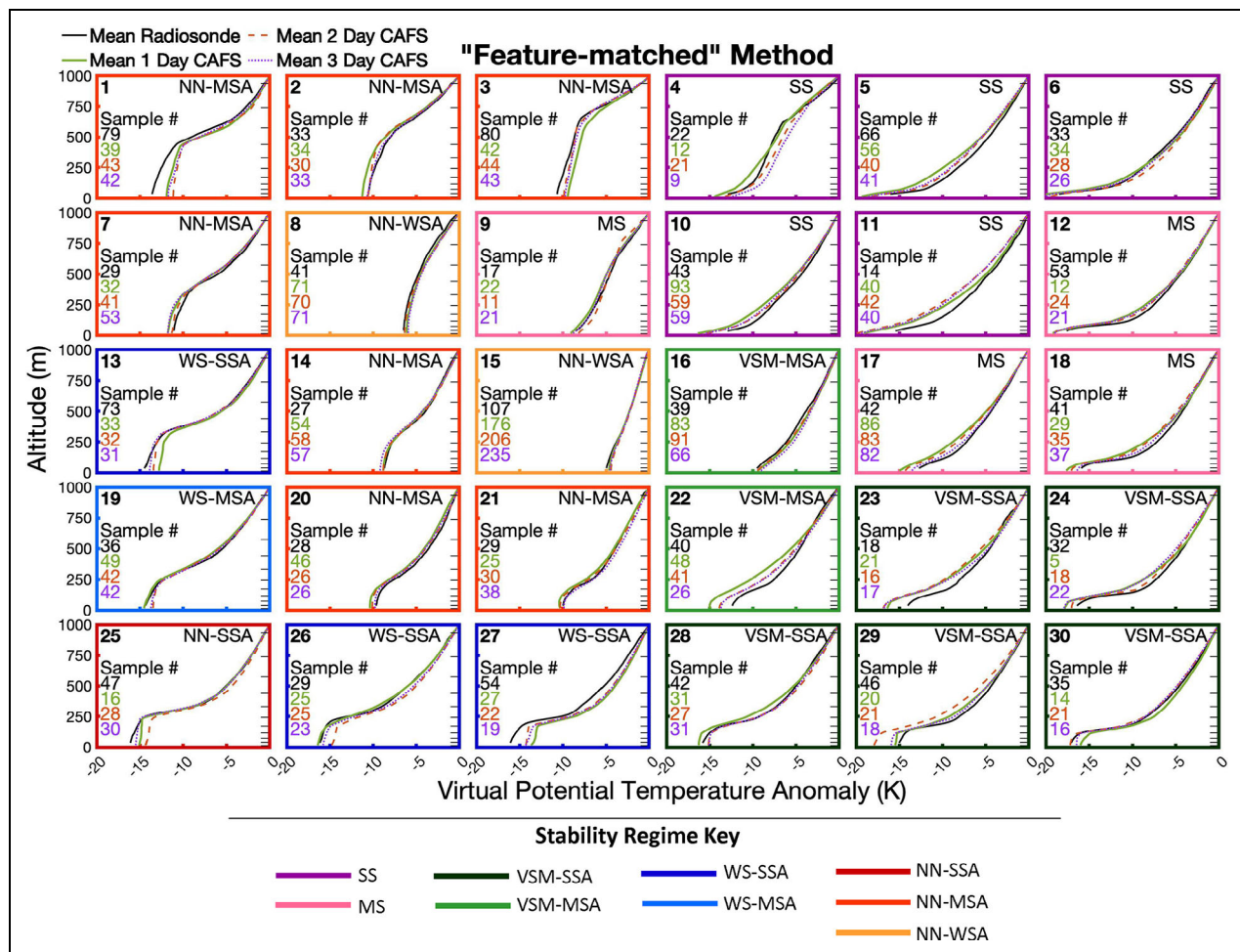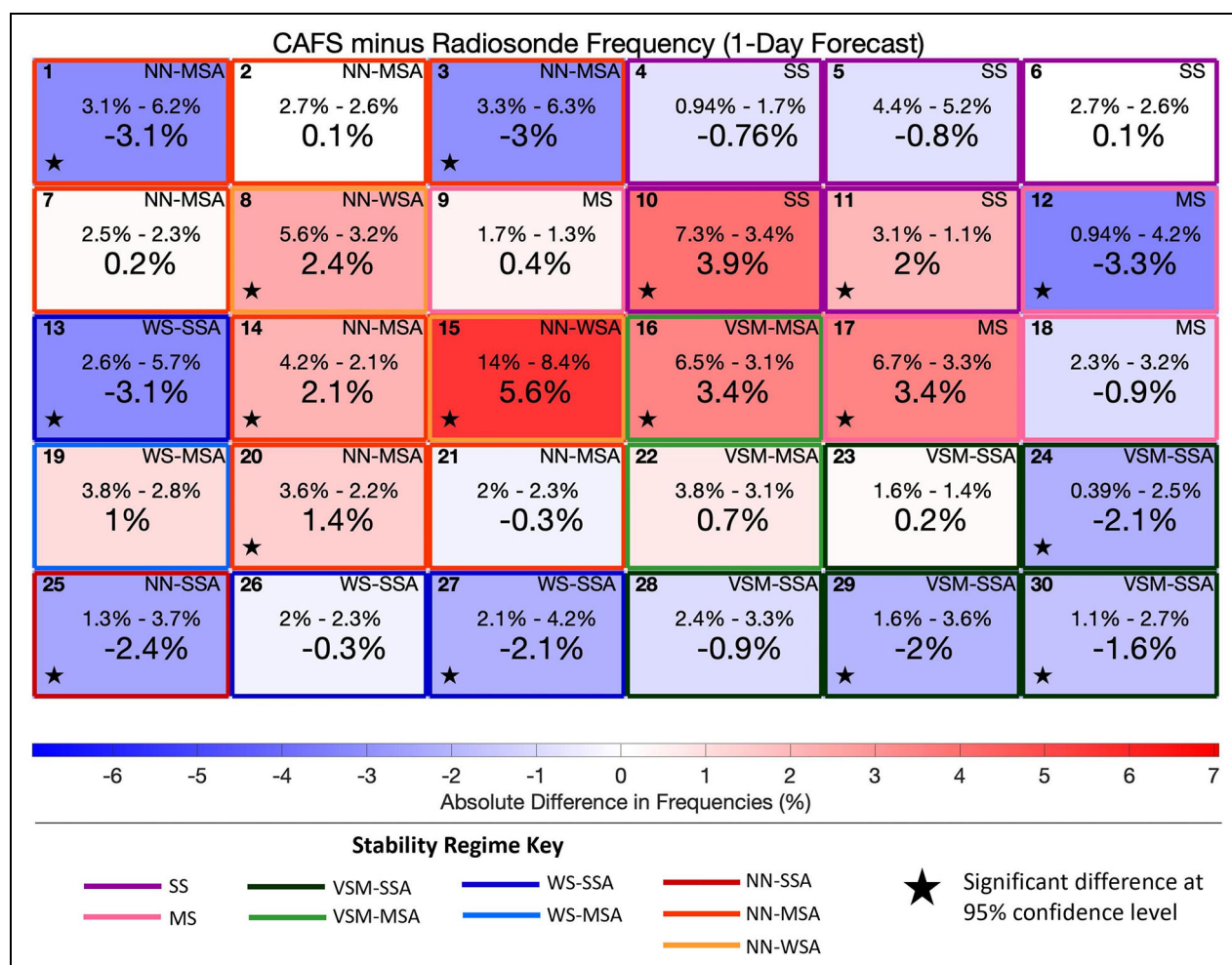
Figure 6. Radiosonde and Coupled Arctic Forecast System (CAFS) profiles assigned to each self-organizing map (SOM) pattern via the "feature-matched" method. Profiles of mean virtual potential temperature (θ_v) anomaly for all radiosonde profiles (black solid) and all CAFS 1-day (green solid), 2-day (orange dashed), and 3-day (purple dotted) forecast profiles. The bold number in the upper left-hand corner of each subplot is the number of that SOM pattern (1 through 30), the numbers on the left side of each subplot are the number of radiosonde (black) and CAFS 1-day (green), 2-day (orange), and 3-day (purple) forecast profiles which map to that pattern, and the letters in the upper right-hand corner of each subplot indicate that pattern's stability regime (**Table 2**). Stability regime is also indicated by the color of the border for each subplot, following the colors given in the "Stability Regime Key." The horizontal tick marks along the right y-axis of each subplot indicate the CAFS native resolution.

forecasts for each SOM pattern, including 25th to 75th percentiles (percentiles from the 2- and 3-day forecasts do not differ discernably from 1-day forecast percentiles). From **Figure 1** (SOM profiles of θ_v anomaly from the "time-matched" method) it was difficult to see whether CAFS simulates all observed θ_v structures, as the CAFS profiles were forced into the various SOM patterns and the mean CAFS profiles washed out many of the finer details of individual profiles. Conversely, for **Figure 6**, CAFS profiles were assigned to the SOM patterns independently from the radiosonde profiles, so only now are we able to see whether CAFS truly simulates all θ_v structures displayed in the SOM.

For patterns with near-surface stability of NN or WS (left half of the SOM), mean CAFS profiles from all forecast periods essentially capture the θ_v inversion strength present in the observation, but for some patterns, CAFS has insufficient stability in the near-surface layer (i.e., the layer is too well-mixed), and capping inversion strength is too

weak. These patterns generally have a lower-level θ_v warm bias with respect to θ_v at 1 km when compared to observations. For patterns with near-surface stability of VSM, MS, or SS (right half of the SOM), mean CAFS profiles from all forecast periods generally capture the proper profile structure, but surface-based (in the case of MS and SS patterns) or elevated (in the case of the VSM patterns) θ_v inversion strength is too great. These patterns generally have a lower-level cold bias with respect to θ_v at 1 km when compared to observations. Overall, results shown in **Figure 6** suggest that CAFS is largely able to reproduce the θ_v profile types observed during MOSAiC, with only minor differences relative to observations.

The total bias and RMSD across the SOM when comparing CAFS to radiosonde profiles (see Supplementary Table S2) decrease with lead time (1-day forecast bias and RMSD are −0.2 K and 0.8 K, respectively, compared to −0.1 K and 0.8 K for the 2-day forecast and −0.1 K and 0.7 K for the 3-

Figure 7. Coupled Arctic Forecast System (CAFS) versus radiosonde self-organizing map (SOM) pattern frequencies. Grid plot following the same layout as the SOM (in **Figure 6**), indicating the absolute difference in frequency between the CAFS 1-day forecast profiles and the radiosonde observations mapped to each SOM pattern. The color bar corresponds to the percent of CAFS cases in each pattern (with respect to all CAFS cases) minus the percent of radiosonde cases in each pattern (with respect to all radiosonde cases), and this absolute difference value is written in the center of each subplot. Above the absolute difference value, the CAFS and radiosonde frequencies are written, separated by a minus sign. A black star indicates a significant difference between the CAFS frequency and the radiosonde frequency. The bold number in the upper left-hand corner of each subplot is the number of that pattern (1 through 30) and the letters in the upper right-hand corner of each subplot indicate that pattern's stability regime (**Table 2**). Stability regime is also indicated by the color of the border for each subplot, following the colors given in the "Stability Regime Key."

day forecast), suggesting that CAFS is better able to recreate correct θ_v profile structures as lead time increases, as CAFS profiles are separately mapped to the SOM for each forecast period. Additionally, bias and RMSD using the "feature-matched" method are an order of magnitude smaller than those calculated using the "time-matched" method.

Thus, we conclude that CAFS is able to recreate observed θ_v profile structures, but not necessarily at the right time (or for the right reasons). This timing inconsistency is demonstrated by the "time-matched" results, where offsets in θ_v inversion height and strength resulted in a smoothing of inversions in the mean CAFS profiles (**Figure 1**).

We also want to know if CAFS simulates the different θ_v profile structures identified in the SOM with the correct frequency. **Figure 7** shows differences in frequency of

each SOM pattern in 1-day CAFS simulations versus in radiosonde observations (CAFS frequency minus radiosonde frequency). Significant differences at the 95% confidence level follow the methods of Cassano et al. (2007). The 2- and 3-day forecasts frequency differences are provided in Supplementary Figure S7. In **Figure 7**, we see that CAFS generally oversimulates patterns in the center of the SOM (red shading; weaker stability throughout the entire profile) and undersimulates patterns around the outside of the SOM (blue shading; stronger stability either near the surface or aloft). The greatest oversimulation is found to occur for pattern 15 (5.6%), which is the most frequently observed pattern. Since the SOM algorithm tends to arrange patterns such that more muted patterns are situated on the inside of the SOM and more extreme
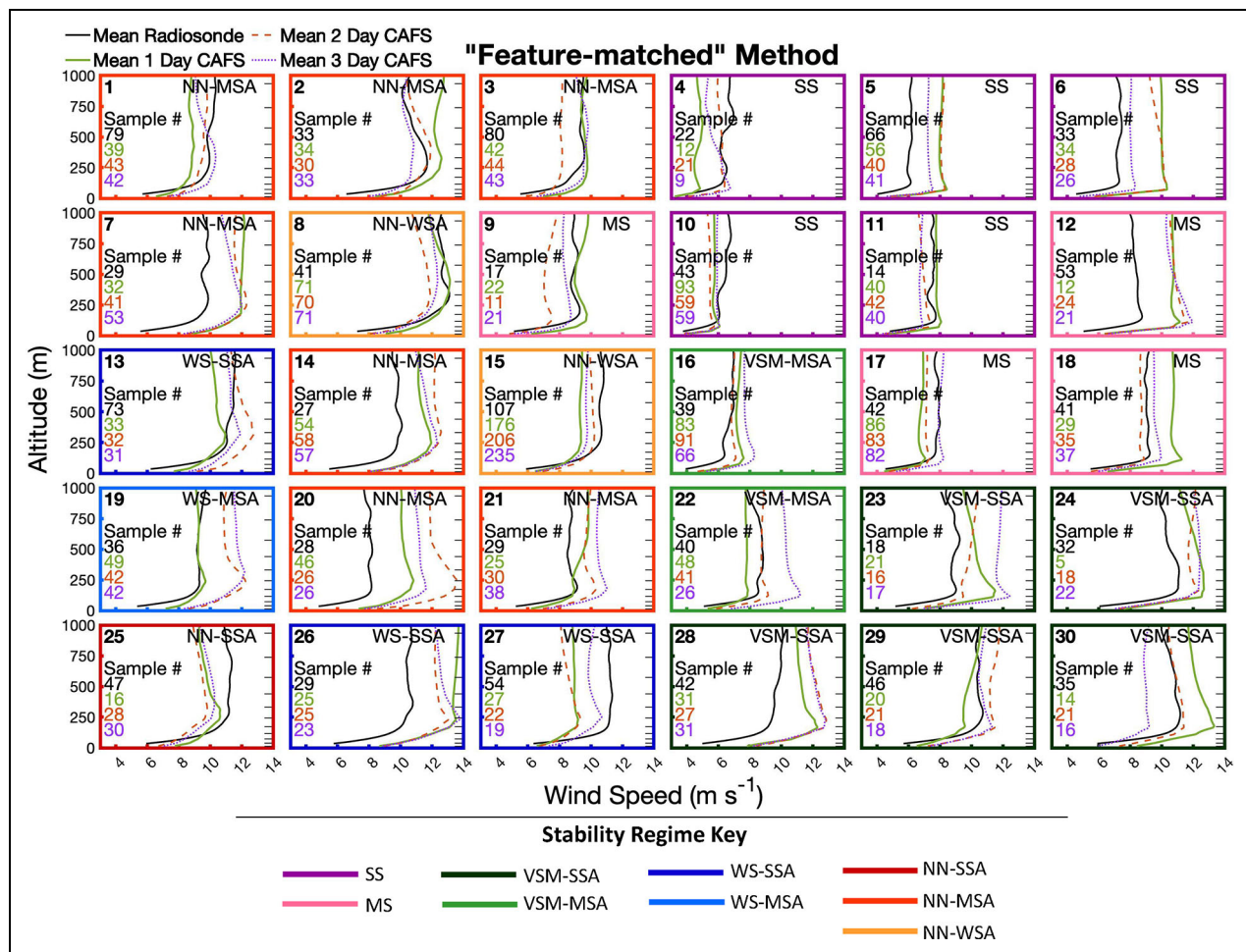
Figure 8. Radiosonde and Coupled Arctic Forecast System (CAFS) profiles assigned to each self-organizing map (SOM) pattern via the "feature-matched" method. As **Figure 6** but for wind speed.

patterns are situated around the outside of the SOM, the results shown in **Figure 7** suggest that CAFS is more easily able to create more muted patterns (thus oversimulating them) and struggles to create more extreme patterns (thus undersimulating them). In this SOM, extreme patterns are those with large θ_v gradients somewhere in the profile. For example, when CAFS simulates SS, it is more likely that the maximum θ_v gradient will be lesser than observed (e.g., compare the θ_v gradients in patterns 10 and 11 to those from patterns 4, 5, and 6 in Jozef et al., 2024).

Mean wind speed profiles from radiosondes and CAFS determined using the "feature-matched" SOM method are shown in **Figure 8**, which reveals whether CAFS simulated the correct wind speed profile structure when the correct θ_v profile structure is simulated. Supplementary Figure S8 shows mean wind speed profiles from radiosondes and CAFS 1-day forecasts for each SOM pattern, including 25th to 75th percentiles (percentiles from the 2- and 3-day forecasts do not differ discernably from 1-day forecast percentiles). For many SOM patterns, mean wind speed profiles differ greatly in shape and magnitude from observed mean wind speed profiles and may be too fast or too slow depending on lead time and height. On the whole, wind speeds in CAFS are too fast, though wind speeds are overall too slow for some SOM patterns (4, 15, 25, and 27). Faster

wind speeds indicate greater wind shear, which would contribute to greater mixing of the near-surface layer. This serves as an explanation as to why CAFS sometimes undersimulates stability below the elevated θ_v inversion seen on the left half of the SOM in **Figure 6** (NN and WS cases). However, this is perhaps counterintuitive for the right half of the SOM (SS, MS, and VSM cases) in which CAFS wind speeds are too fast but stability is too strong for some patterns since stronger winds are associated with greater turbulent mixing and weaker stability (Jozef et al. 2023a). This suggests that CAFS simulates greater decoupling between the near-surface layer and the overlying atmosphere than actually occurs, such that stronger stability can persist even in high wind speed environments.

Interestingly, for many SOM patterns, despite simulating wind speed profiles that differ from those observed, CAFS is still able to simulate correct θ_v profile structure (**Figure 6**). Thus, suggests that the relationship between wind speed and stability in CAFS is incorrect, there is another forcing mechanism in CAFS that dominates over the incorrect wind speed to still produce the correct θ_v structure, or some combination of both reasonings. This is further supported by the fact that for a given stability regime, CAFS overestimates wind speed for some SOM patterns but underestimates it for others (e.g., pattern 26 vs. pattern 27),
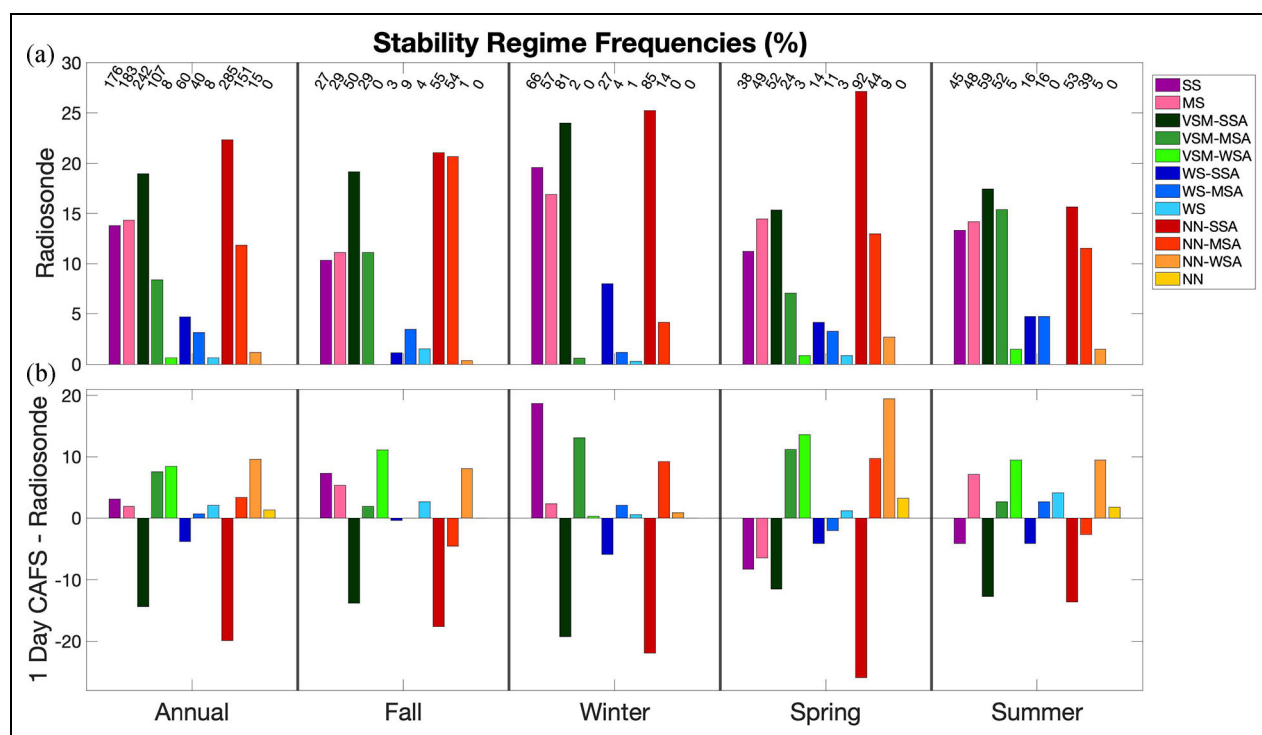
Figure 9. Coupled Arctic Forecast System (CAFS) versus radiosonde stability regime frequencies. (a) Frequency distribution showing the percent of radiosonde profiles in each stability regime (**Table 2**), annually and seasonally. For seasonal sections, the percent shown is with respect to the total number of radiosonde profiles in that season. The numbers along the top of the plot, above each bar, indicate the total number of radiosonde profiles of that stability regime and season. (b) Absolute difference in frequency between the CAFS 1-day forecast profiles and radiosonde observations in each stability regime, annually and seasonally. A horizonal line at zero is included for reference.

suggesting that θ_v profile structure in CAFS can be minimally impacted by a change in wind speed.

Additionally, it varies for each SOM pattern whether the 1-, 2-, or 3-day forecasts produce wind speed profiles most similar to the observations, and total bias and RMSD when comparing CAFS to radiosonde profiles for the entire SOM (see Supplementary Table S2) are similar regardless of lead time (1-day forecast bias and RMSD are 0.6 m s^{-1} and 1.6 m s^{-1}, respectively, compared to 0.6 m s^{-1} and 1.7 m s^{-1} for the 2-day forecast and 0.8 m s^{-1} and 1.6 m s^{-1} for the 3-day forecast). Thus, simulation of the wind speed vertical structure corresponding to a given θ_v profile structure does not appear to improve or degrade with lead time.

3.2.2. Stability regime analysis

In this section, we delve into whether CAFS is simulating stability regimes with the correct frequencies, and how the mechanical and radiative forcings for a given stability regime differ from those observed. We additionally break the analysis into fall (Sep–Nov), winter (Dec–Feb), spring (Mar–May), and summer (Jun–Aug) to reveal any seasonal dependencies.

First, we show (**Figure 9**) the observed frequency distribution of stability regimes, as well as the differences in the frequency of occurrence of the various regimes as simulated by CAFS and as observed (CAFS frequency minus radiosonde frequency) for the 1-day forecast (frequency differences for the 2- and 3-day forecasts are shown in Supplementary Figure S9). It must be noted that the number of radiosonde

cases of a certain stability regime in **Figure 9** may not match the total number of cases mapped to SOM patterns of that stability regime (e.g., as shown in **Figure 6**). This is because cases are mapped to the SOM pattern with overall the most similar profile shape, but which may ultimately be categorized as different stability regime. Thus not all individual cases will have the same stability regime as that SOM pattern to which they are mapped.

Here, a negative difference in frequency indicates that the CAFS-simulated frequency of a particular stability regime was lower than observed, and a positive difference indicates that the CAFS-simulated frequency was higher than observed.

The most common stability regimes in the observations (**Figure 9a**) are VSM-SSA and NN-SSA, the frequencies of which are both greatly underestimated by CAFS annually (by 14% and 20%, respectively) and in all individual seasons (**Figure 9b**). WS-SSA is not observed as frequently but is also underrepresented by CAFS (by 3.8%). Near-surface stability of VSM, WS, or NN with moderate or weak stability aloft (-MSA or -WSA) is observed less frequently than strong stability aloft (especially -WSA), and all such stability regimes are overrepresented by CAFS annually (0.7%–7.5% for -MSA and 2.1%–9.6% for -WSA). This is true seasonally as well, except for NN-MSA in fall and summer and WS-MSA in spring, which are underrepresented. Additionally, annually and in spring and summer, the frequency of NN (which was never observed) is
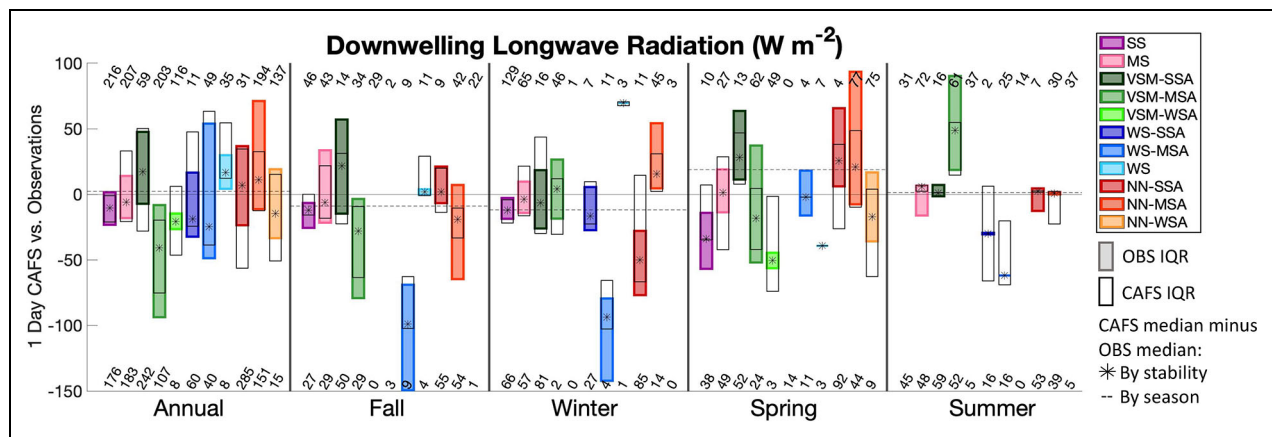
Figure 10. Coupled Arctic Forecast System (CAFS) versus observed downwelling longwave radiation (LWD). Difference in observed LWD and that simulated by the CAFS 1-day forecast depending on stability regime (**Table 2**), annually and seasonally. For each stability regime, the center asterisk indicates the CAFS median value minus the observed (OBS) median value. The boxes indicate the interquartile range (IQR) of the observational (filled) and CAFS (open) data, presented as the spread around the difference between the CAFS and observed median (asterisk). Dashed horizontal black lines indicate the difference in median from all CAFS data and all observations annually and for each season, and a solid horizontal black line at zero is included for reference. The numbers along the top and bottom indicate the number of CAFS profiles and radiosonde profiles of that stability regime and season, respectively.

overestimated by CAFS (by 1.3% annually and by 3.2% and 1.8% in spring and summer, respectively). SS and MS are the next most frequently observed regimes aside from VSM-SSA and NN-SSA (**Figure 9a**), both of which are overrepresented by CAFS annually (by 3.1% and 1.9%, respectively). When broken into seasons, SS and MS are both overrepresented in fall and winter (the greatest overestimation is for SS in winter, of 19%) and underrepresented in spring. In summer, SS is underrepresented while MS is overrepresented. The difference in stability frequencies when compared with observations in 2- and 3-day forecasts are largely similar to those in 1-day forecasts. However one notable difference is that the annual overestimation of SS and MS decrease (to 0% and 0.8%, respectively, for the 2-day forecast) and ultimately become an underestimation (of 0.9% and 1.1%, respectively, for the 3-day forecast) as lead time increases.

Thus, on the whole, CAFS struggles to simulate strong stability aloft and overrepresents the frequency of a strongly or moderately stable near-surface atmosphere. One exception to the previous statement is that both strong and moderate near-surface stability are underrepresented in spring, and SS is underrepresented in summer. Looking at profiles of the bias in θ_v anomaly between CAFS and the radiosonde observations separated by season (Supplementary Figure S10), it is evident that bias is greatest in spring (followed by summer) among all the seasons (mean bias of 3 K in spring vs. 0.9 K annually for 1-day forecasts). This enhanced positive θ_v bias in CAFS near the surface (i.e., a decreased temperature gradient) indicates that CAFS is undersimulating stability, particularly in spring. Therefore, we expect to see greater differences in radiative and mechanical forcings between CAFS and the observations in spring versus the other seasons.

Through the "feature-mapped" stability regime analyses, we can assess if CAFS simulates a similar range of LWD

values across the different stability regimes as is observed. This is depicted in **Figure 10**, which shows differences in LWD in 1-day CAFS forecasts compared to observed LWD for each stability regime (results using 2- and 3-day CAFS forecasts are shown in Supplementary Figure S11). In **Figure 10**, differences between CAFS-simulated and observed LWD are shown for each stability regime and season, where asterisks represent the median of LWD from CAFS minus that from the observations for each stability regime, such that a value above zero indicates that CAFS overestimates LWD, and a value below zero indicates that CAFS underestimates LWD. Next, boxes surrounding the asterisk indicate the interquartile range (IQR) of the data using the difference in medians as a reference point, such that the top of the box is the value at the asterisk plus the difference between the CAFS (observation) median and the maximum of the CAFS (observation) IQR and the bottom of the box is the value at the asterisk minus the difference between the CAFS (observation) median and the minimum of the CAFS (observation) IQR. The overall difference in median LWD annually and for each season between CAFS and the observations are shown in **Figure 10** and are presented in Supplementary Table S3.

Annually, the overall difference in median LWD between CAFS and observations shows that CAFS has a small positive bias in LWD of 2.2 W m^{-2}. This positive bias is seen for the individual regimes of VSM-SSA (17 W m^{-2}), WS (17 W m^{-2}), NN-SSA (6.8 W m^{-2}), and NN-WSA (11 W m^{-2}). There is a negative bias in LWD with a magnitude less than 25 W m^{-2} for all other regimes, except VSM-MSA which has a bias of −41 W m^{-2}. In general, the strongest (SS and MS) and weakest (NN) stability regimes have biases closest to zero, revealing that CAFS better represents the relationship between stability and LWD on the 2 ends of the stability spectrum, but struggles for the regimes in the middle (VSM and WS).
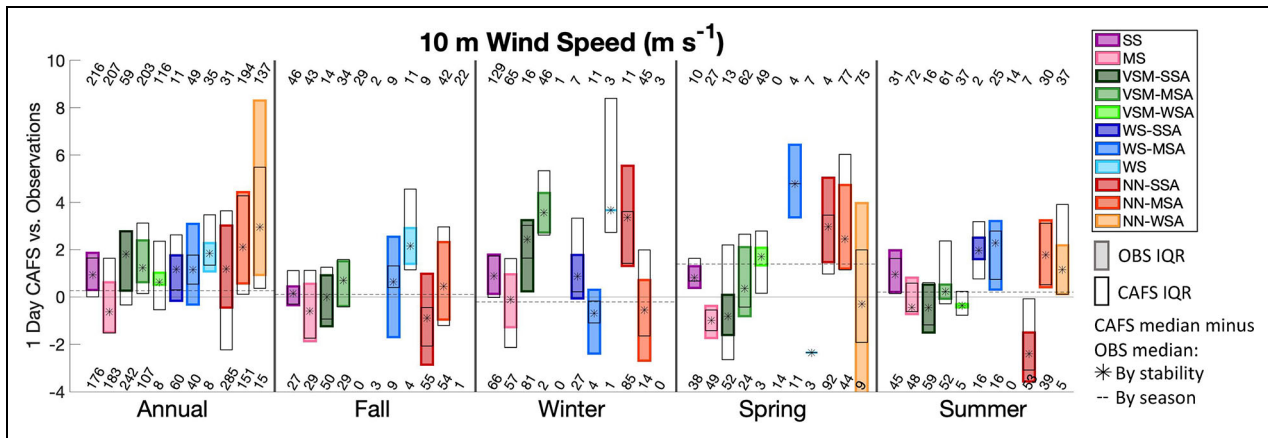
Figure 11. Coupled Arctic Forecast System (CAFS) versus observed 10 m wind speed. As **Figure 10** but for 10 m wind speed.

Within individual seasons, patterns described in the previous paragraph are most similar to those observed in spring when relationships between stability regime and median difference generally follow those seen annually, although the seasonal bias in LWD is larger than that seen annually at 19 W m^{-2}. As increased LWD contributes to relatively warmer near-surface temperatures, this explains why there is a large positive θ_v bias near the surface in spring, while other seasons show a lesser positive or a negative θ_v bias (Supplementary Figure S10), and why weaker stability is more overrepresented in spring than in other seasons (**Figure 9**). The results during spring being most similar to the annual results is consistent with the findings of Jozef et al. (2023a) which showed the relationships between turbulent forcing mechanisms and stability regime observed during spring to match the annual patterns more closely than any other season. The reason for this is likely because atmospheric conditions during spring are highly variable and can be reminiscent of common conditions during each of the other seasons.

In fall and winter on a whole, CAFS underrepresents LWD (bias of -9 W m^{-2} and -12 W m^{-2}, respectively), which is also true for many individual stability regimes. The difference between simulated and observed LWD is closer to zero for most individual regimes in fall and winter versus annually, especially VSM in winter. In summer, the CAFS LWD is most similar to that of the observations (bias of 1.4 W m^{-2}). Thus, CAFS better simulates total LWD in fall, winter, and summer, compared to spring. This variability in difference between simulated and observed LWD to create certain stability regimes across seasons suggests CAFS is not properly simulating the seasonal dependence of the relationship between stability and LWD, specifically manifesting in CAFS simulating excessive LWD values in spring but not in the other seasons.

Some stability regimes show similar annual distributions of LWD in CAFS and the observations. However, the distributions are quite different for MS, VSM-SSA, VSM-WSA, WS-SSA, WS, NN-SSA, and NN-WSA, in which CAFS shows a greater spread than was observed. This indicates that CAFS has a wider range of LWD conditions, which might result in each of these regimes. Differences in

distributions from CAFS and the observations are greatest in summer. Since summer also has the lowest overall bias in LWD, it becomes clear that CAFS struggles to simulate the complex relationship between LWD and stability during summer months (as discussed in Jozef et al., 2023a).

Results for the 2- and 3-day CAFS forecasts are similar to those from the 1-day forecast. However, there are some differences. For example, the overall difference in medians between CAFS and observations is positive for all seasons except fall in the 2- and 3-day forecasts. Thus, overestimation of LWD is enhanced with greater lead time. This finding suggests that there may be a greater occurrence of model errors resulting in higher LWD (such as the over-simulation of clouds, or simulation of clouds with excessive liquid water content) for longer forecast times. This is consistent with an overrepresentation of SS (of about 3% for the 1-day forecast) shifting to an underrepresentation of SS (of about 1% for the 3-day forecast) with greater lead time (Supplementary Figure S9).

Through the "feature-mapped" stability regime analyses, we can also assess how CAFS 10 m wind speed accuracy varies with stability regime. **Figure 11** shows how forcing from 10 m wind speed in 1-day CAFS forecasts differs from that observed by the met tower (results using the 2- and 3-day CAFS forecasts are shown in Supplementary Figure S12). Overall annual and seasonal differences between CAFS and observational medians are presented in Supplementary Table S4.

Annually, the overall difference in median 10 m wind speeds shows that CAFS has a positive bias of 0.3 m s^{-1}, and this positive bias holds true for all individual stability regimes except MS. The greatest difference in median 10 m wind speeds occurs for NN-WSA (3.0 m s^{-1}) and NN-MSA (2.1 m s^{-1}), and the smallest difference in median occurs for VSM-WSA (0.6 m s^{-1}) and SS (0.9 m s^{-1}), suggesting that CAFS better simulates wind speed in cases with near-surface strong stability or a VSM layer versus cases with NN stability. Overestimation of 10 m wind speeds is almost ubiquitous throughout all seasons (aside from an overall bias of -0.2 m s^{-1} in winter, and a negative bias for some individual regimes throughout the seasons), with the greatest bias in spring (1.4 m s^{-1}).

On the whole, positive biases increase with decreasing stability. This suggests that an increase in wind speed in CAFS has a greater impact on stability (i.e., causes greater weakening) than observed. However for some combinations of season and near-surface stability of NN, the positive bias is lower than for stronger stability regimes, or is negative (e.g., NN-WSA in spring). The magnitudes of biases across seasons are largely similar to the annual biases for the various stability regimes (this is especially true for the stronger stability regimes), which suggests a general consistency in the relationship between simulated 10 m wind speed and stability regime throughout the year.

When comparing the IQRs annually, the distribution of 10 m wind speed simulated in CAFS is wider than that in the observations for all regimes except WS-MSA and NN-WSA. For regimes with stronger stability near the surface and aloft (SS, MS, and VSM-SSA), the distributions from CAFS and the observations are most similar. This again suggests that for weaker stability regimes, mechanical forcing in CAFS is less comparable to the observed forcing than for more stable regimes.

Results for the 2- and 3-day CAFS forecasts are similar to those from the 1-day forecast, with some exceptions. For example, overall differences in median simulated and observed wind speeds are positive for all seasons in the 2- and 3-day forecasts. Thus, overestimation of 10 m wind speed is enhanced with greater lead time (annual overestimation of 0.3 m s^{-1} for the 1-day forecast vs. 0.5 m s^{-1} for both the 2- and 3-day forecasts). As this trend was also seen for LWD, this suggests that model errors develop such that both greater radiative and mechanical forcings are present (such as stormy conditions that bring both clouds and wind) with increased lead time.

4. Summary and conclusions
In this study, we assessed the ability of the CAFS to simulate the lower atmospheric vertical structure and stability as observed during the MOSAiC expedition. By assigning CAFS profiles to a SOM via a "time-matched" method (assigning a CAFS profile to the SOM pattern of its corresponding radiosonde observation; **Figure 1**), it was revealed that CAFS does not necessarily simulate the right θ_v vertical structure at the right time. However, by assigning CAFS profiles to a SOM via a "feature-matched" method (assigning a CAFS profile to a SOM pattern by minimizing the squared difference between the two; **Figure 6**), it was revealed that CAFS can simulate the range of θ_v vertical structures observed (albeit at the wrong times), where more muted patterns (those with a weaker θ_v gradient) are overrepresented by CAFS and more extreme patterns (those with a stronger θ_v gradient) are underrepresented by CAFS (**Figure 7**). For example, when CAFS simulates SS, the near-surface θ_v gradient (while still within the valid range for SS) is likely to be weaker than frequently observed.

The vertical structures of wind speed on the SOM match the observations well for the "time-matched" comparisons (**Figure 2**), but not for the "feature-matched" comparisons (**Figure 8**), suggesting that CAFS does

simulate the right wind speeds at the right time, but the relationship between wind speed and θ_v vertical structure in CAFS does not match that observed. This suggests that the relative impacts of various turbulent forcing mechanisms on θ_v vertical structure, including wind speed, are not simulated properly. This largely manifests in excessive winds in CAFS contributing to a certain observed θ_v structure. Additionally, excessive wind speeds in CAFS sometimes correspond to even stronger stability in CAFS than observed, suggesting that CAFS oversimulates decoupling between a stable ABL and the overlying atmosphere. This phenomenon is most prevalent in winter (**Figure 11**).

When determining stability regimes for CAFS and observational profiles, it was discovered that CAFS underestimates the frequency of strong stability aloft in all seasons, as well as strong stability near the surface in spring and summer. Conversely, CAFS overestimates the frequency of moderate and weak stability aloft, as well as near-surface strong stability in fall and winter (**Figure 9**). CAFS is best at simulating SS and NN-WSA at the right times, but when CAFS stability is wrong, it is usually too weak rather than too stable (**Figure 3**).

To understand how differences in radiative and mechanical forcings corresponding to stability regimes in CAFS versus in the observations contribute to the discrepancies in stability structure and frequency, we compared LWD (**Figures 4** and **10**) and 10 m wind speed (**Figures 5** and **11**) in CAFS to observations. When considering all values, LWD and 10 m wind speed are slightly overestimated in CAFS, but CAFS is generally representing these variables well since the overestimations are very small. One exception is during spring, when the biases in both wind speed and LWD are very high (1.4 m s^{-1} and 19 W m^{-2}, respectively). When considering individual stability regimes, 10 m wind speed is nearly ubiquitously overestimated, and overestimation increases with decreasing stability. However, the sign of the bias in LWD depends on stability regime, with negative biases for the SS and WS regimes (regimes with generally lower observed LWD values), and positive biases for the VSM and NN regimes (regimes with generally higher observed LWD values). Thus, for SS and the WS regimes, there are opposing biases in the 2 turbulent forcing mechanisms (positive wind speed bias and negative LWD bias). Therefore, these regimes are likely occurring in CAFS because the negative LWD bias, which favors stronger stability, is offset by the positive wind speed bias, which tends to weaken stability. However for cases where both mechanical (wind speed) and thermodynamic (LWD) forcings are overestimated in CAFS, the ABL will become well-mixed (either as VSM or NN depending on the magnitude of the forcings).

When stability in CAFS is weaker (stronger) than observed, biases in the forcing are positive (negative). One exception is that LWD has a negative bias when VSM was observed and WS near-surface stability was simulated and vice versa. All of the aforementioned results, in combination, explain why CAFS overestimates (underestimates) the frequency of weak (strong) stability aloft, especially in spring. Overall, analysis of radiative and mechanical forcings in CAFS reveals that the model does generate the

proper correlation between turbulent forcings and stability (i.e., higher amounts of LWD and faster wind speeds contribute to weaker stability, though VSM cases correspond to higher amounts of LWD than WS). However, the magnitude of these forcings for a given stability regime is inexact. Future work should focus on discovering why CAFS is simulating incorrect magnitudes of the 10 m wind speed and LWD for a given stability regime.

The results of this analysis help to document and illustrate both strengths (e.g., ability to create the correct vertical θ_v structures and generally reflecting the correct correlations between turbulent forcings and stability) and weaknesses (e.g., inability to always produce the right θ_v structures at the right times or with the correct frequencies, and a consistent offset in the magnitudes of turbulent forcings which produce the various stability regimes) in the CAFS model. Future work should expand upon these findings to explore additional atmospheric features which may be affecting the efficacy of CAFS (e.g., inconsistencies in simulated cloud cover vs. observed may explain many of the differences found in this article). Another direction for future work could be to assess the efficacy of CAFS during transition events, such as during a warm air intrusion or the passing of a storm, through a case study analysis comparing observations from the DataHawk2 uncrewed aircraft system (which have the added benefit of turbulence measurements and a lower starting altitude; de Boer et al., 2022), radiosondes, and other airborne instruments (e.g., tethered balloon) to CAFS output. Additionally, the approach used to evaluate CAFS in this article will be applied to a multimodel ensemble based on Arctic CORDEX (Coordinated Regional Downscaling Experiment) simulations.

Data accessibility statement
The level 2 radiosonde data used in this study are available at the PANGAEA Data Publisher at https://doi.org/10.1594/PANGAEA.928656 (Maturilli et al., 2021). Meteorological tower data and radiation are available at the National Science Foundation Arctic Data Center at https://doi.org/10.18739/A2PV6B83F (Cox et al., 2023a) as described in Cox et al. (2023b). CAFS data are available at the PANGAEA Data Publisher at https://doi.org/10.1594/PANGAEA.964776 (Solomon and Jozef, 2024).

Supplemental files
The supplemental files for this article can be found as follows:

Supplementary Material.pdf

Acknowledgments

Funding

Competing interests
The authors have declared that no competing interests exist.

Author contributions
Conceptualized the analysis presented in this article: GJ, JJC, GdB.

Provided the CAFS data: AS.

Provided insight on the CAFS model: AS, JI.

Analyzed the data: GJ.

Wrote the manuscript: GJ.

Reviewed and edited the manuscript: GJ, JJC, GdB, AS, JI.

Approved the submitted version for publication: GJ, JJC, GdB, AS, JI.

References
Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung. 2017. Polar research and supply vessel POLARSTERN operated by the Alfred-Wegener-Institute. *Journal of Large-Scale Research Facilities* **3**: A119. DOI: http://dx.doi.org/10.17815/jlsrf-3-163.

Andreas, EL, **Horst, TW**, **Grachev, AA**, **Persson, POG**, **Fairall, CW**, **Guest, PS**, **Jordan, RE.** 2010. Parametrizing turbulent exchange over summer sea ice and the marginal ice zone. *Quarterly Journal of the Royal Meteorological Society* **136**(649): 927–943. DOI: http://dx.doi.org/10.1002/qj.618.

Baklanov, AA, **Grisogono, B**, **Bornstein, R**, **Mahrt, L**, **Zilitinkevich, SS**, **Taylor, P**, **Larsen, SE**, **Rotach, MW**, **Fernando, HJS.** 2011. The nature, theory, and modeling of atmospheric planetary boundary layers. *Bulletin of the American Meteorological Society* **92**(2): 123–128. DOI: http://dx.doi.org/10.1175/2010BAMS2797.1.

Banta, RM. 2008. Stable-boundary-layer regimes from the perspective of the low-level jet. *Acta Geophysica* **56**: 58–87. DOI: http://dx.doi.org/10.2478/s11600-007-0049-8.

Banta, RM, **Mahrt, L**, **Vickers, D**, **Sun, J**, **Balsley, BB**, **Pichugina, YL**, **Williams, EJ.** 2007. The very stable boundary layer on nights with weak low-level jets. *Journal of the Atmospheric Sciences* **64**(9): 3068–3090. DOI: http://dx.doi.org/10.1175/JAS4002.1.

Banta, RM, **Pechugina, YL**, **Newsom, RK.** 2003. Relationship between low-level jet properties and turbulence kinetic energy in the nocturnal stable boundary layer. *Journal of the Atmospheric Sciences* **60**(20): 2549–2555. DOI: http://dx.doi.org/10.1175/1520-0469(2003)060<2549:RBLJPA>2.0.CO;2.

Berry, DI, **Moat, BI**, **Yelland, MJ.** 2001. Airflow distortion at instrument sites on the FS Polarstern. Southampton, UK: Southampton Oceanography Centre. Southampton Oceanography Centre Internal Document 69: 39 pp. Available at https://nora.nerc.ac.uk/id/eprint/502825.

Brooks, IM, **Tjernström, M**, **Persson, POG**, **Shupe, MD**, **Atkinson, RA**, **Canut, G**, **Birch, CE**, **Mauritsen, T**, **Sedlar, J**, **Brooks, BJ.** 2017. The turbulent structure of the Arctic summer boundary layer during the Arctic summer cloud-ocean study. *Journal of Geophysical Research: Atmospheres* **122**(18): 9685–9704. DOI: http://dx.doi.org/10.1002/2017JD027234.

Cassano, EN, **Glisan, JM**, **Cassano, JJ**, **Gutowski, WJ Jr**, **Seefeldt, MW.** 2015. Self-organizing map analysis of widespread temperature extremes in Alaska and Canada. *Climate Research* **62**: 199–218. DOI: http://dx.doi.org/10.3354/cr01274.

Cassano, EN, **Lynch, AH**, **Cassano, JJ**, **Koslow, MR.** 2006. Classification of synoptic patterns in the western Arctic associated with extreme events at Barrow, Alaska, USA. *Climate Research* **30**: 83–97. DOI: http://dx.doi.org/10.3354/cr030083.

Cassano, JJ, **Nigro, M**, **Lazzara, M.** 2016. Characteristics of the near-surface atmosphere over the Ross Ice Shelf, Antarctica. *Journal of Geophysical Research: Atmospheres* **121**(7): 3339–3362. DOI: http://dx.doi.org/10.1002/2015JD024383.

Cassano, JJ, **Uotila, P**, **Lynch, AH**, **Cassano, EN.** 2007. Predicted changes in synoptic forcing of net precipitation in large Arctic river basins during the 21st century. *Journal of Geophysical Research* **112**(G4): G04S49. DOI: http://dx.doi.org/10.1029/2006JG000332.

Chechin, DG, **Lüpkes, C**, **Hartmann, J**, **Ehrlich, A**, **Wendisch, M.** 2023. Turbulent structure of the Arctic boundary layer in early summer driven by stability, wind shear and cloud-top radiative cooling: ACLOUD airborne observations. *Atmospheric Chemistry and Physics* **23**(8): 4685–4707. DOI: http://dx.doi.org/10.5194/acp-23-4685-2023.

Cohen, J, **Zhang, X**, **Francis, J**, **Jung, T**, **Kwok, R**, **Overland, J**, **Ballinger, TJ**, **Bhatt, US**, **Chen, HW**, **Coumou, D**, **Feldstein, S**, **Gu, H**, **Handorf, D**, **Henderson, G**, **Ionita, M**, **Kretschmer, M**, **Laliberte, F**, **Lee, S**, **Linderholm, HW**, **Maslowski, W**, **Peings, Y**, **Pfeiffer, K**, **Rigor, I**, **Semmler, T**, **Stroeve, J**, **Taylor, PC**, **Vavrus, S**, **Vihma, T**, **Wang, S**, **Wendisch, M**, **Wu, Y**, **Yoon, J.** 2020. Divergent consensuses on Arctic amplification influence on mid-latitude severe winter weather. *Nature Climate Change* **10**: 20–29. DOI: http://dx.doi.org/10.1038/s41558-019-0662-y.

Cox, C, **Gallagher, M**, **Shupe, M**, **Persson, O**, **Blomquist, B**, **Grachev, A**, **Riihimaki, L**, **Kutchenreiter, M**, **Morris, V**, **Solomon, A**, **Brooks, I**, **Costa, D**, **Gottas, D**, **Hutchings, J**, **Osborn, J**, **Morris, S**, **Preusser, A**, **Uttal, T.** 2023a. Met city meteorological and surface flux measurements (Level 3 Final), Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC), Central Arctic, October 2019–September 2020. Arctic Data Center [dataset]. DOI: http://dx.doi.org/10.18739/A2PV6B83F.

Cox, CJ, **Gallagher, M**, **Shupe, MD**, **Persson, POG**, **Solomon, A**, **Fairall, CW**, **Ayers, T**, **Blomquist, B**, **Brooks, IM**, **Costa, D**, **Grachev, A**, **Gottas, D**, **Hutchings, JK**, **Kutchenreiter, M**, **Leach, J**, **Morris, SM**, **Morris, V**, **Osborn, J**, **Pezoa, S**, **Preusser, A**, **Riihimaki, L**, **Uttal, T.** 2023b. Continuous observations of the surface energy budget and meteorology over the Arctic sea ice during MOSAiC. *Scientific Data* **10**: 519. DOI: http://dx.doi.org/10.1038/s41597-023-02415-5.

Cuxart, J, **Holtslag, AAM**, **Beare, RJ**, **Bazile, E**, **Beljaars, A**, **Cheng, A**, **Conangla, L**, **Ek, M**, **Freedman, F**, **Hamdi, R**, **Kerstein, A**, **Kitagawa, H**, **Lenderink, G**, **Lewellen, D**, **Mailhot, J**, **Mauritsen, T**, **Perov, V**, **Schayes, G**, **Steeneveld, G-J**, **Svensson, G**, **Taylor, P**, **Weng, W**, **Wunsch, S**, **Xu, K-M.** 2006. Single-column model intercomparison for a stably stratified atmospheric boundary layer. *Boundary-Layer Meteorology* **118**: 273–303. DOI: http://dx.doi.org/10.1007/s10546-005-3780-1.

de Boer, G, **Calmer, R**, **Jozef, G**, **Cassano, J**, **Hamilton, J**, **Lawrence, D**, **Borenstein, S**, **Doddi, A**, **Cox, C**, **Schmale, J**, **Preußer, A**, **Argrow, B.** 2022. Observing the Central Arctic atmosphere and surface with University of Colorado uncrewed aircraft systems. *Scientific Data* **9**: 439. DOI: http://dx.doi.org/10.1038/s41597-022-01526-9.

Deser, C, **Tomas, RA**, **Sun, L.** 2015. The role of ocean–atmosphere coupling in the zonal-mean atmospheric response to Arctic sea ice loss. *Journal of Climate* **28**(6): 2168–2186. DOI: http://dx.doi.org/10.1175/JCLI-D-14-00325.1.

Dice, MJ, **Cassano, JJ.** 2022. Assessing physical relationships between atmospheric state, fluxes, and boundary layer stability at McMurdo Station, Antarctica. *Journal of Geophysical Research: Atmospheres* **127**(15): e2021JD036075. DOI: http://dx.doi.org/10.1029/2021JD036075.

Dice, MJ, **Cassano, JJ**, **Jozef, GC.** 2024. Forcing for varying boundary layer stability across Antarctica. *Weather and Climate Dynamics* **5**(1): 369–394. DOI: http://dx.doi.org/10.5194/wcd-5-369-2024.

Dice, MJ, **Cassano, JJ**, **Jozef, GC**, **Seefeldt, M.** 2023. Variations in boundary layer stability across Antarctica: A comparison between coastal and interior

sites. *Weather and Climate Dynamics* **4**(4): 1045–1069. DOI: http://dx.doi.org/10.5194/wcd-4-1045-2023.

Esau, I, Sorokina, S. 2010. Climatology of the Arctic planetary boundary layer, in Lang, PR, Lombargo, FS eds., *Atmospheirc turbulence, meteorological modeling and aerodynamics*. New York, NY: Nova Science Publishers Inc: 3–58.

Francis, JA, Vavrus, SJ. 2015. Evidence for a wavier jet stream in response to rapid Arctic warming. *Environmental Research Letters* **10**: 014005. DOI: http://dx.doi.org/10.1088/1748-9326/10/1/014005.

Hamilton, J, de Boer, G, Doddi, A, Lawrence, DA. 2022. The DataHawk2 uncrewed aircraft system for atmospheric research. *Atmospheric Measurement Techniques* **15**(22): 6789–6806. DOI: http://dx.doi.org/10.5194/amt-15-6789-2022.

Holtslag, AAM, Svensson, G, Baas, P, Basu, S, Beare, B, Beljaars, ACM, Bosveld, FC, Cuxart, J, Lindvall, J, Steeneveld, G-J, Tjernström, M, Van De Wiel, BJH. 2013. Stable atmospheric boundary layers and diurnal cycles: Challenges for weather and climate models. *Bulletin of the American Meteorological Society* **94**(11): 1691–1706. DOI: http://dx.doi.org/10.1175/BAMS-D-11-00187.1.

Huang, Y, Ding, Q, Dong, X, Xi, B, Baxter, I. 2021. Summertime low clouds mediate the impact of the large-scale circulation on Arctic sea ice. *Communications Earth & Environment* **2**(1): 38. DOI: http://dx.doi.org/10.1038/s43247-021-00114-w.

Jenkins, M, Dai, A. 2021. The impact of sea-ice loss on Arctic climate feedbacks and their role for Arctic amplification. *Geophysical Research Letters* **48**(15): e2021GL094599. DOI: http://dx.doi.org/10.1029/2021GL094599.

Jozef, G, Cassano, J, Dahlke, S, de Boer, G. 2022. Testing the efficacy of atmospheric boundary layer height detection algorithms using uncrewed aircraft system data from MOSAiC. *Atmospheric Measurement Techniques* **15**(13): 4001–4022. DOI: http://dx.doi.org/10.5194/amt-15-4001-2022.

Jozef, GC, Cassano, JJ, Dahlke, S, Dice, M, Cox, CJ, de Boer, G. 2023a. Thermodynamic and kinematic drivers of atmospheric boundary layer stability in the central Arctic during the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC). *Atmospheric Chemistry and Physics* **23**(20): 13087–13106. DOI: http://dx.doi.org/10.5194/acp-23-13087-2023.

Jozef, GC, Cassano, JJ, Dahlke, S, Dice, M, Cox, CJ, de Boer, G. 2024. An overview of the vertical structure of the atmospheric boundary layer in the central Arctic during MOSAiC. *Atmospheric Chemistry and Physics* **24**(2): 1429–1450. DOI: http://dx.doi.org/10.5194/acp-24-1429-2024.

Jozef, GC, Klingel, R, Cassano, JJ, Maronga, B, de Boer, G, Dahlke, S, Cox, CJ. 2023b. Derivation and compilation of lower-atmospheric properties relating to temperature, wind, stability, moisture, and surface radiation budget over the central Arctic sea ice

during MOSAiC. *Earth System Science Data* **15**(11): 4983–4995. DOI: http://dx.doi.org/10.5194/essd-15-4983-2023.

Kohonen, T. 2001. *Self-organizing maps. 3rd ed.* Berlin, Germany: Springer. DOI: http://dx.doi.org/10.1007/978-3-642-56927-2.

Labe, Z, Peings, Y, Magnusdottir, G. 2020. Warm Arctic, cold Siberia pattern: Role of full Arctic amplification versus sea ice loss alone. *Geophysical Research Letters* **47**(17): e2020GL088583. DOI: http://dx.doi.org/10.1029/2020GL088583.

Li, X, Mace, GG, Strong, C, Krueger, SK. 2023. Wintertime cooling of the Arctic TOA by low-level clouds. *Geophysical Research Letters* **50**(17): e2023GL104869. DOI: http://dx.doi.org/10.1029/2023GL104869.

Liu, Y, Weisburg, RH. 2011. A review of self-organizing map applications in meteorology and oceanography, in Mwasiagi, JI ed., *Self organizing maps—Applications and novel algorithm design*. Rijeka, Croatia: InTech: 253–268.

Lüpkes, C, Vihma, T, Bimbaum, G, Wacker, U. 2008. Influence of leads in sea ice on the temperature of the atmospheric boundary layer during polar night. *Geophysical Research Letters* **35**(8): 2–6. DOI: http://dx.doi.org/10.1029/2007GL032461.

Marsik, FJ, Fischer, KW, McDonald, TD, Samson, PJ. 1995. Comparison of methods for estimating mixing height used during the 1992 Atlanta field intensive. *Journal of Applied Meteorology and Climatology* **34**(8): 1802–1814. DOI: http://dx.doi.org/10.1175/1520-0450(1995)034<1802:COMFEM>2.0.CO;2.

Maslowski, W, Kinney, JC, Higging, M, Roberts, A. 2012. The future of Arctic Sea ice. *Annual Review of Earth and Planetary Sciences.* **40**: 625–654. DOI: http://dx.doi.org/10.1146/annurev-earth-042711-105345.

Maturilli, M, Holdridge, DJ, Dahlke, S, Graeser, J, Sommerfeld, A, Jaiser, R, Deckelmann, H, Schulz, A. 2021. Initial radiosonde data from 2019-10 to 2020-09 during project MOSAiC. Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, PANGAEA [dataset]. DOI: http://dx.doi.org/10.1594/PANGAEA.928656.

Maturilli, M, Sommer, M, Holdridge, DJ, Dahlke, S, Graeser, J, Sommerfeld, A, Jaiser, R, Deckelmann, H, Schulz, A. 2022. MOSAiC radiosonde data (level 3). PANGAEA [dataset]. DOI: http://dx.doi.org/10.1594/PANGAEA.943870.

Melia, N, Haines, K, Hawkins, E. 2016. Sea ice decline and 21st century trans-Arctic shipping routes. *Geophysical Research Letters* **43**(18): 9720–9728. DOI: http://dx.doi.org/10.1002/2016GL069315.

National Centers for Environmental Information. 2020. Global Forecast System (GFS). Available at https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast. Accessed April 25, 2023.

Nigro, MA, Cassano, JJ, Willi, J, Bromwich, DH, Lazzara, MA. 2017. A self-organizing-map-based evaluation of the Antarctic mesoscale prediction system using observations from a 30-m instrumented tower

on the Ross Ice Shelf, Antarctica. *Weather and Forecasting* **32**(1): 223–242. DOI: http://dx.doi.org/10.1175/WAF-D-16-0084.1.

Nixdorf, U, Dethloff, K, Rex, M, Shupe, M, Sommerfeld, A, Perovich, D, Nicolaus, M, Heuzé, C, Rabe, B, Loose, B, Damm, E, Gradinger, R, Fong, A, Maslowski, W, Rinke, A, Kwok, R, Spreen, G, Wendisch, M, Herber, A, Hirsekorn, M, Mohaupt, V, Frickenhaus, S, Immerz, A, Weiss-Tuider, K, König, B, Mengedoht, D, Regnery, J, Gerchow, P, Ransby, D, Krumpen, T, Morgenstern, A, Haas, C, Kanzow, T, Rack, FR, Saitzev, V, Sokolov, V, Makarov, A, Schwarze, S, Wunderlich, T, Wurr, K, Boetius, A. 2021. MOSAiC Extended acknowledgement. Zenodo. DOI: http://dx.doi.org/10.5281/zenodo.5179738.

NOAA Physical Sciences Laboratory. n.d. What is CAFS? Available at https://psl.noaa.gov/forecasts/seaice/about.html. Accessed April 25, 2023.

Persson, POG, Fairall, CW, Andreas, EL, Guest, PS, Perovich, DK. 2002. Measurements near the Atmospheric Surface Flux Group tower at SHEBA: Near-surface conditions and surface energy budget. *Journal of Geophysical Research: Oceans* **107**(C10): SHE 21-1–SHE 21-35. DOI: http://dx.doi.org/10.1029/2000JC000705.

Pithan, F, Ackerman, A, Angevine, WM, Hartung, K, Ickes, L, Kelley, M, Medeiros, B, Sandu, I, Steeneveld, G-J, Sterk, HAM, Svenssen, G, Vaillancourt, PA, Zadra, A. 2016. Select strengths and biases of models in representing the Arctic winter boundary layer over sea ice: The Larcform 1 single column model intercomparison. *Journal of Advances in Modeling Earth Systems* **8**(3): 1345–1357. DOI: http://dx.doi.org/10.1002/2016MS000630.

Rantanen, M, Karpechko, AY, Lipponen, A, Nordling, K, Hyvärinen, O, Ruosteenoja, K, Vihma, T, Laarksonen, A. 2022. The Arctic has warmed nearly four times faster than the globe since 1979. *Communications Earth & Environment* **3**: 168. DOI: http://dx.doi.org/10.1038/s43247-022-00498-3.

Schneider, T, Lüpkes, C, Dorn, W, Chechin, D, Handorf, D, Khosravi, S, Gryanik, VM, Makhotina, I, Rinke, A. 2021. Sensitivity to changes in the surface-layer turbulence parameterization for stable conditions in winter: A case study with a regional climate model over the Arctic. *Atmospheric Science Letters* **23**(1): e1066. DOI: http://dx.doi.org/10.1002/asl.1066.

Screen, JA. 2021. An ice-free Arctic: What could it mean for European weather? *Weather* **76**(10): 327–328. DOI: http://dx.doi.org/10.1002/wea.4069.

Serreze, MC, Barry, RG. 2011. Processes and impacts of Arctic amplification: A research synthesis. *Global and Planetary Change* **77**(1–2): 85–96. DOI: http://dx.doi.org/10.1016/j.gloplacha.2011.03.004.

Serreze, MC, Francis, JA. 2006. The Arctic amplification debate. *Climatic Change* **76**: 241–264. DOI: http://dx.doi.org/10.1007/s10584-005-9017-y.

Sheridan, SC, Lee, CC. 2011. The self-organizing map in synoptic climatological research. *Progress in Physical Geography: Earth and Environment* **35**(1): 109–119. DOI: http://dx.doi.org/10.1177/0309133310397582.

Shupe, MD, Rex, M, Blomquist, B, Persson, POG, Schmale, J, Uttal, T, Althausen, D, Angot, H, Archer, S, Bariteau, L, Beck, I, Bilberry, J, Bucci, 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, Lonardi, M, Loose, B, 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, ZQ, Yue, F. 2022. Overview of the MOSAiC expedition: Atmosphere. *Elementa: Science of the Anthropocene* **10**: 00060. DOI: http://dx.doi.org/10.1525/elementa.2021.00060.

Shupe, MD, Rex, M, Dethloff, K, Damm, E, Fong, AA, Gradinger, R, Heuzé, C, Loose, B, Makarov, A, Maslowski, W, Nicolaus, M, Perovich, D, Rabe, B, Rinke, A, Sokolov, V, Sommerfeld, A. 2020. The MOSAiC expedition: A year drifting with the Arctic sea ice. NOAA Arctic Report Card, National Oceanic and Atmospheric Administration (NOAA): 1–8. DOI: http://dx.doi.org/10.25923/9g3v-xh92.

Solomon, A, Jozef, G. 2024. Coupled Arctic Forecast System (CAFS) forecasts used in study evaluating atmospheric stability with MOSAiC soundings. PANGAEA [dataset]. DOI: http://dx.doi.org/10.1594/PANGAEA.964776.

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): 00104. DOI: http://dx.doi.org/10.1525/elementa.2022.00104.

Steeneveld, G-J, Van De Wiel, BJH, Holtslag, AAM. 2006. Modelling the Arctic stable boundary layer and its coupling to the surface. *Boundary-Layer Meteorology* **118**: 357–378. DOI: http://dx.doi.org/10.1007/s10546-005-7771-z.

Stroeve, JC, **Kattsov, V**, **Barrett, A**, **Serreze, M**, **Pavlova, T**, **Holland, M**, **Meier, WN.** 2012. Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations. *Geophysical Research Letters* **39**(16): L16502. DOI: http://dx.doi.org/10.1029/2012GL052676.

Stull, RB. 1988. *An introduction to boundary layer meteorology.* Amsterdam, the Netherlands: Kluwer Academic Publishers. DOI: http://dx.doi.org/10.1007/978-94-009-3027-8.

Tjernström, M, **Graversen, RG.** 2009. The vertical structure of the lower Arctic troposphere analysed from observations and the ERA-40 reanalysis. *Quarterly Journal of the Royal Meteorological Society* **135**(639): 431–443. DOI: http://dx.doi.org/10.1002/qj.380.

Tjernström, M, **Leck, C**, **Persson, POG**, **Jensen, ML**, **Oncley, SP**, **Targino, A.** 2004. The summertime Arctic atmosphere: Meteorological measurements during the Arctic Ocean experiment 2001. *Bulletin of the American Meteorological Society* **85**(9): 1305–1321. DOI: http://dx.doi.org/10.1175/BAMS-85-9-1305.

Tjernström, M, **Zagar, M**, **Svensson, G**, **Cassano, JJ**, **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**: 337–381. DOI: http://dx.doi.org/10.1007/s10546-004-7954-z.

Tomas, RA, **Deser, C**, **Sun, L.** 2016. The role of ocean heat transport in the global climate response to projected Arctic sea ice loss. *Journal of Climate* **29**(19): 6841–6859. DOI: http://dx.doi.org/10.1175/JCLI-D-15-0651.1.

Vihma, T, **Lüpkes, C**, **Hartmann, J**, **Savijärvi, H.** 2005. Observations and modelling of cold-air advection over Arctic sea ice. *Boundary-Layer Meteorology* **117**: 275–300. DOI: http://dx.doi.org/10.1007/s10546-004-6005-0.

Wang, S, **Wang, Q**, **Jordan, RE**, **Persson, POG.** 2001. Interactions among longwave radiation of clouds, turbulence, and snow surface temperature in the Arctic: A model sensitivity study. *Journal of Geophysical Research* **106**(D14): 15323–15333. DOI: http://dx.doi.org/10.1029/2000JD900358.