A machine learning correction model of the winter clear-sky temperature bias over the Arctic sea ice in atmospheric reanalyses

Lorenzo Zampieri[1], Gabriele Arduini[2], Marika Holland[1], Sarah Keeley[2], Kristian S. Mogensen[2], Matthew D. Shupe[3,4], and Steffen Tietsche[2]

[1]National Center for Atmospheric Research
[2]European Centre for Medium-Range Weather Forecasts
[3]Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, CO, USA
[4]National Oceanic and Atmospheric Administration, Physical Science Laboratory, Boulder, CO, USA

February 27, 2023

A machine learning correction model of the winter clear-sky temperature bias over the Arctic sea ice in atmospheric reanalyses

Lorenzo Zampieri[a] , Gabriele Arduini[b] , Marika Holland[a] , Sarah Keeley[b] , Kristian Mogensen[b] , Matthew D. Shupe[c,d] , Steffen Tietsche[b]

[a] *National Center for Atmospheric Research, Boulder, CO, USA ,*

[b] *European Centre for Medium-range Weather Forecasts, Reading, United Kingdom and Bonn, Germany,*

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

[d] *National Oceanic and Atmospheric Administration, Physical Science Laboratory, Boulder, CO, USA*

Corresponding author: Lorenzo Zampieri, zampieri@ucar.edu

ABSTRACT: Atmospheric reanalyses are widely used to estimate the past atmospheric near-surface state over sea ice. They provide boundary conditions for sea ice and ocean numerical simulations and relevant information for studying polar variability and anthropogenic climate change. Previous research revealed the existence of large near-surface temperature biases (mostly warm) over the Arctic sea ice in the current generation of atmospheric reanalyses, which is linked to a poor representation of the snow over the sea ice and the stably stratified boundary layer in the forecast models used to produce the reanalyses. These errors can compromise the employment of reanalysis products in support of polar research. Here, we train a fully connected neural network that learns from remote sensing infrared temperature observations to correct the existing generation of uncoupled atmospheric reanalyses (ERA5, JRA-55) based on a set of sea ice and atmospheric predictors, which are themselves reanalysis products. The advantages of the proposed correction scheme over previous calibration attempts are the consideration of the synoptic weather and cloud state, compatibility of the predictors with the mechanism responsible for the bias, and a self-emerging seasonality and multi-decadal trend consistent with the declining sea ice state in the Arctic. The correction leads on average to a 27% temperature bias reduction for ERA5 and 7% for JRA-55 if compared to independent in-situ observations from the MOSAiC campaign (respectively 32% and 10% under clear-sky conditions). These improvements can be beneficial for forced sea ice and ocean simulations, which rely on reanalyses surface fields as boundary conditions.

SIGNIFICANCE STATEMENT: This study illustrates a novel method based on machine learning for reducing the systematic surface temperature errors that characterize multiple atmospheric reanalyses in sea-ice-covered regions of the Arctic under clear-sky conditions. The correction applied to the temperature field is consistent with the local weather and the sea ice and snow conditions, meaning that it responds to seasonal changes in sea ice cover as well as to its long-term decline due to global warming. The corrected reanalysis temperature can be employed to support polar research activities, and in particular to better simulate the evolution of the interacting sea ice and ocean system within numerical models.

Copyright information

1. Introduction

An atmospheric reanalysis is a realistic retrospective description of the atmospheric state obtained by constraining an atmospheric model simulation with observations through the application of data assimilation techniques. The resulting products are continuously available over a relatively long period (currently the last 40 to 70 years), retain consistency because they are realized with a single model and data assimilation version, and feature a uniform and continuous spatial coverage (Lindsay et al. 2014). This is a particularly desirable property in the polar regions, where only a few in-situ environmental observations are available (Jung et al. 2016). For these reasons, reanalyses are widely used as an estimate for the present and past atmospheric near-surface state over the Arctic sea ice, with one relevant application being to serve as boundary conditions for sea ice and ocean simulations (Large and Yeager 2008; Tsujino et al. 2018), fundamental tools to study the effects of climate change on the polar regions and to predict the sea-ice evolution at various timescales.

Because of the lack of measurements assimilated over the polar regions by the reanalysis models, the near-surface Arctic atmospheric state is only weakly constrained by observations and strongly dependent on the formulation of the models, and this can lead to errors when this formulation is not appropriate (Zampieri et al. 2018, 2019). Furthermore, when measurements are available, the presence of a shallow atmospheric boundary layer and temperature inversion—challenging features to simulate correctly even for state-of-the-art models—reduces the effectiveness of the assimilation procedure. In this respect, previous research revealed large surface temperature biases over the Arctic sea ice for most atmospheric reanalyses (Tjernström and Graversen 2009), a fact that has been later linked to a poor representation of the snow and sea-ice state in the numerical surface schemes of the reanalysis models (Batrak and Müller 2019). Most reanalysis models prescribe a constant sea ice thickness in time and space and do not account for the presence of a snow layer over the sea ice, erroneously quantifying the insulating effect of the sea ice system and thus the heat conduction through this medium. As a result, the reanalyses surface temperature tends to be too warm in regions where the real insulating effect of ice and snow would be larger than that prescribed in the models, and too cold in regions where the sea ice and snow are thin and consequently exhibit lower insulating properties (Fig. 3 of Batrak and Müller (2019)). Given the intra- and inter-annual spatiotemporal variability of the sea ice and snow thickness in the Arctic, the resulting model biases tend to be heterogeneous but particularly accentuated during winter Clear Sky Events (CSE), when the surface experiences strong radiative cooling (Serreze et al. 2007), a process hard to simulate correctly without modeling the insulating snow layer over the sea ice.

Numerical Weather Prediction (NWP) centers will likely address this model deficiency in future reanalysis versions by employing fully coupled modelling systems (Keeley and Mogensen 2018; Arduini et al. 2022; Day et al. 2022) and assimilating new kinds of near-surface observations. A first step in this direction has been taken in the C3S Arctic Regional Reanalysis (Copernicus Climate Change Service 2021), where the snow over sea ice is modeled more accurately. Nevertheless, the reduction of the temperature bias in coupled systems is still subordinated to a correct simulation of the sea ice system, and in particular the snow and sea ice thickness. Meanwhile, this study explores the possibility of correcting offline the existing generation of uncoupled reanalyses by training a Machine Learning (ML) algorithm that links key atmospheric and sea ice variables to a realistic estimate of the surface temperature carefully derived from remote sensing surface observations that are currently not assimilated in the reanalyses models. The resulting correction is by design state-dependent and therefore consistent with the large-scale Arctic weather, as well as the declining trend of the sea ice thickness. Furthermore, it increases the heterogeneity and realism of the reanalysis surface state in sea ice regions, and it can be derived seamlessly in time and space because it relies entirely on reanalysis-based predictors. Our correction model can be adapted to multiple reanalysis products but here we focus in particular on the European Centre for Medium-range Weather Forecasts (ECMWF) Reanalysis version 5 (Hersbach et al. 2020) (ERA5) and the Japanese Meteorological Agency second reanalysis project (Onogi et al. 2007; Kobayashi et al. 2015) (JRA-55), arguably among the most used reanalyses for sea ice and polar applications. The main objectives of this study are summarized in the following points:

1. Presenting the methodology behind the ML bias correction strategy for the skin surface temperature over sea ice, including its practical implementation.
2. Quantifying the bias reduction and describing the relation of the correction with the sea ice and atmospheric states.
3. Analyzing the seasonality and interannual variability of the correction, including its impact on the historical warming trend observed in the Arctic during recent years.

2. Methods

This section provides details on the ML algorithm used to correct the atmospheric reanalysis, the datasets employed for its training and validation, and the criteria for its application. The reader should note that, in practice, two identical correction models are trained and employed in parallel for this study, one for each reanalysis product considered. Unless otherwise stated, these ML models share the same network structure (but different weights estimates) and therefore the description in the method section will be generalized to keep the exposition more compact and clear. Prior to presenting the correction strategy, we begin with a description of the observations that serve as an improved estimate of the surface temperature and have key implications for the correction model itself.

a. Satellite Observations of the Ice Surface Temperature

While typically not a problem when investigating slow evolving sea ice variables such as the sea ice concentration, the sub-daily variability of the temperature field can be substantial due to the evolution of the local weather and changes in insolation. For these reasons, this quantity can vary at the sub-daily timescales in both observations and reanalyses even if polar regions experience a reduced or absent daily cycle for most of the year. This study employs swath-based temperature observations, commonly referred to as Level 2, to capture this sub-daily temperature variability. More information on the data levels definitions can be found at `https://www.earthdata.nasa.gov/engage/open-data-services-and-software/data-information-policy/data-levels`. A Level 2 product type informs us of the exact time and location a satellite observation was taken.

The swath-based satellite data used in this study are from the Arctic and Antarctic Ice Surface Temperatures from thermal Infrared satellite sensors dataset (AASTI; Høyer et al. (2019)), available from 2000 to 2009. This dataset is based on the work of Høyer and She (2007); Høyer et al. (2014); Rasmussen et al. (2018) at the Danish Meteorological Institute and it was created in the framework of the EUSTACE project (EU Surface Temperature for All Corners of Earth). The dataset is built by combining observations from the Advanced Very High Resolution Radiometer (AVHRR) instruments onboard different satellites of the National Oceanic and Atmospheric Administration (NOAA) and the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT; see Fig. 2 in Nielsen-Englyst et al. (2021) for further details on the observational platforms). Only clear-sky observations are included in the dataset and considered for this study. In cloudy-sky conditions, the satellite sensor would measure the thermal signature of the cloud top rather than that of the sea ice or snow at the surface. The total uncertainty of the AASTI observations is on the order of $\sim 2^o C$. The uncertainty is partitioned into three components: random uncertainty, locally systematic uncertainty, and large-scale systematic uncertainty (Nielsen-Englyst et al. 2021). A quality level flag from 1 (bad data) to 5 (best quality) is provided, and in this study, we consider only observations with quality levels 3, 4, and 5. The observations have a spatial resolution of $\sim 0.05^o$, meaning that they can resolve the temperature signal of ice features with a typical length scale of a few kilometers, such as big leads, coastal polynyas, and extensive sea ice floes. Because the Arctic sea surface is characterized by the occurrence of open water and newly refrozen leads down to the meter scale (Thielke et al. 2022), there can be a certain level of ambiguity regarding what surface type is represented by the temperature observation. This additional source of uncertainty cannot be easily taken into account: the temperature retrieval algorithm is nonlinear, and the exact ice surface temperature cannot be reconstructed based on the observed sea ice concentration. However, this aspect does not affect our study substantially, as we focus on the winter season and the pack-ice regions, which feature the occurrence of open water only sporadically mostly due to dynamical sea ice processes.

Finally, the reader should note that in Fig. 1c, we show the daily aggregated number of surface temperature observations from a Level 3 dataset (Dybkjær et al. 2012) rather than the Level 2 AASTI dataset used to train the correction model.

b. The Machine Learning Bias Correction Model

Network Predictors

As already mentioned in Sec 1, previous studies have highlighted links between the reanalyses temperature bias and different aspects of the atmosphere and sea ice systems, such as the cloud state, the sea ice and snow thickness, and the surface atmospheric temperature itself. Based on the previous considerations, the following four model predictors have been chosen as input for the ML model:

SKT Reanalysis Skin Temperature: The skin temperature is the theoretical temperature that is required to satisfy the surface energy balance. This temperature is converted to an ice-only temperature based on the reanalyses open water fraction. This is the same field we aim to ultimately correct.

STRD Reanalysis Surface Downward Longwave Radiation: This physical quantity is the amount of thermal (or longwave) radiation emitted by the atmosphere and clouds that reaches a horizontal plane at the surface.

SIT Sea Ice Thickness: The sea ice thickness represents the average depth of sea ice observed inside a grid cell. Here, we do not use in-situ thickness measurements or remote sensing retrievals of this quantity due a high fragmentation in time and space. Instead, a gap-free reanalysis-based estimate from the Pan-Arctic Ice Ocean Modeling and Assimilation System (PIOMAS) (Zhang and Rothrock 2003) is obtained by dividing the point-wise volume of sea ice per unit area by the sea ice area fraction.

SND Snow Thickness on Sea Ice: Similarly to the sea ice thickness, the snow thickness estimates employed here also come from a reanalysis product, the SnowModel-LG (Liston et al. 2018, 2020), where a Lagrangian snow-evolution model forced with the precipitation from the

ERA5 atmospheric reanalysis is used to produce daily pan-Arctic snow-on-sea-ice depth distributions.

The predictors can be divided into an atmospheric group (SKT and STRD), and in an ice group (SIT and SND). The source of SKT and STRD changes according to the atmospheric reanalysis product under consideration, while SIT and SND remain the same for all reanalyses. The output data used to train the network is defined as the difference between the original reanalysis skin temperature and the surface temperature observations described in Sec. 2a. To build the training dataset for the ML correction model, all the input variables are interpolated to the exact location and time of the observations by using a bi-linear interpolation scheme provided by the Xarray Python package (Hoyer and Hamman 2017). Being all model-based reanalysis fields, the inputs are available over the whole Arctic domain for 40 years (01.08.1980 to 31.07.2021), allowing the temperature correction to be consistently computed over sea ice regions without spatiotemporal gaps if observations were available to fully characterize the bias. Because the snow and sea ice thickness data are not available for some isolated ocean points along the coastlines due to grid conversion issues, we filled these points with data from the nearest neighboring grid cells. This occurrence is rare and confined to complex coastal domains (e.g. the Canadian Archipelago). Ultimately, the resulting temperature correction has the same time-step as the atmospheric predictors SKT and STRD (1h for ERA5, 3h for JRA-55).

A further correction skill source could come from the inclusion of the wind speed among the predictors. Based on our physical intuition, the turbulent heat flux tends to decrease in low-wind conditions, enhancing the radiative cooling and the boundary layer stratification. On the contrary, in high-wind conditions the heat is redistributed much more efficiently between the surface and the boundary layer, reducing the importance of the ice state in determining the surface temperature. At present, this aspect is outside the scope of our work and therefore not considered in the current manuscript, but we acknowledge the potential of a better representation of the turbulence and stratification in our model design.

Network Design

A fully connected neural network (NN) has been chosen to model the reanalysis temperature correction because it is flexible, easy to implement and train, and able to capture the nonlinear relations between the system state and the correction. After testing different network designs, we chose a simple setup consisting of a Deep Feed Forward (DFF) NN with 5 hidden layers featuring 16 nodes each, resulting in 80 trainable weights. All the network nodes, except those linearly activated belonging to the last layer, feature a standard "ReLu" activation function. The network cost function is minimized using an "Adam" algorithm, a mean squared error loss function is employed, and the learning rate is 0.01. Note that the uncertainties of the observations are not taken into account during the minimization process of the cost function. The chosen batch size is 1024 and the training epochs are 10. The correction model was developed in Python based on the Pytorch package (Paszke et al. 2019).

The network inputs have been normalized with a linear transformation to fit the interval $[-1;+1]$. This ML standard procedure is necessary since the NN input data combines different physical quantities with values spanning several orders of magnitude. This fact could induce the NN to overweight some predictors while neglecting others. The size of the NN combined dataset varies depending on the reanalysis in consideration because of the different spatiotemporal resolutions, but it remains in the order of 5×10^7 points collected over the period 01.2000–12.2009 for both ERA5 and JRA-55. The data are divided into training, validation, and test subsets following a simple approach that guarantees that neighboring data points, which are likely correlated, are not distributed into more than one subset. First, we subdivide the dataset into multiple five-day portions. For each of these, the first three days are dedicated to the training subset, the fourth day to the validation subset, and the fifth day to the testing subset. The three subsets are then shuffled separately before the training step. The test subset provides an unbiased evaluation of the final model fit on the dataset by using data never seen by the model during the training and validation phase. All the plots presented in the next section of this paper refer to the test subset. The training and validation phases of the correction model were completed in approximately one wall-clock hour when run on a single cluster node with 72 processors.

c. Application Criteria of the Bias Correction Model

Given the features of observations and reanalyses presented in the previous paragraphs, we conclude that the correction model should not be applied indiscriminately to the entire Arctic domain but rather to the regions experiencing clear-sky conditions, where observations are more

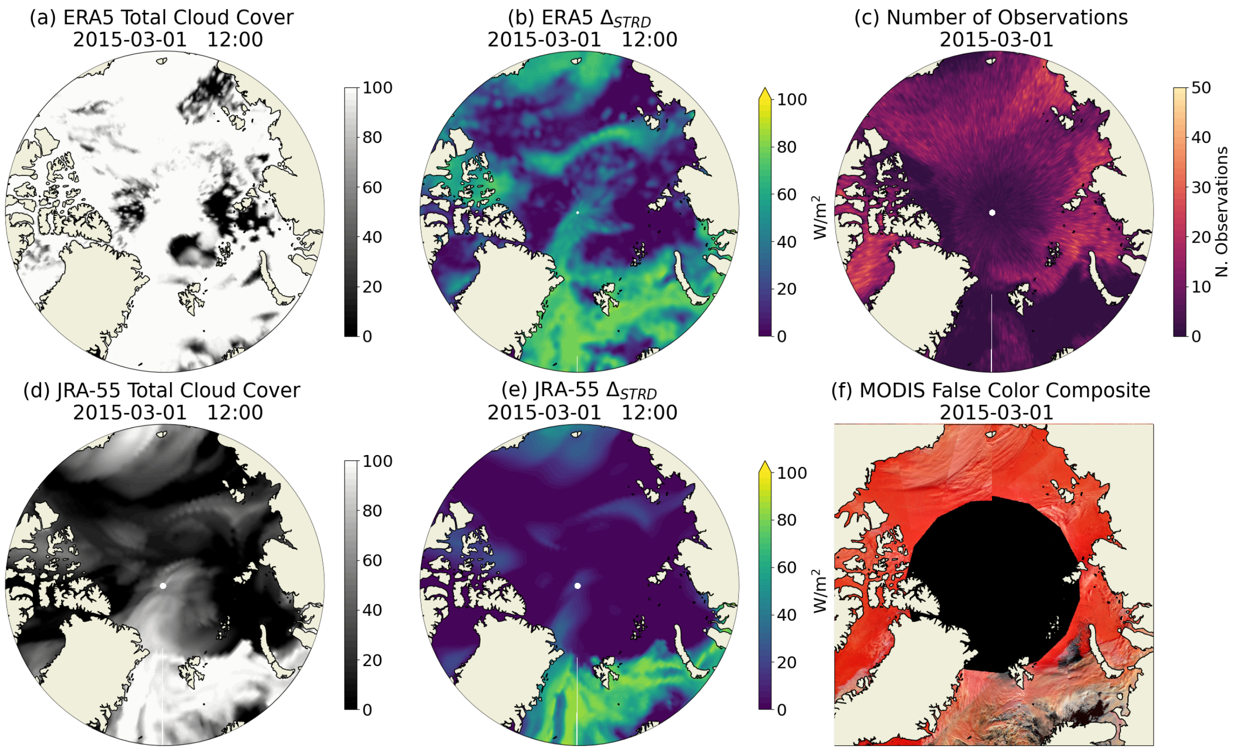

Fig. 1. **(a)** ERA5 total cloud coverage (TCC) on 2015-03-01 at 12:00. **(b)** Difference between the ERA5 all-sky and clear-sky surface downward thermal radiation on 2015-03-01 at 12:00 (Δ_{STRD}). Low values of Δ_{STRD} are an indication of little or absent cloud coverage. **(c)** Number of observations collected by the AVHRR satellite sensors orbiting on 2015-03-01. An high observation count is an indication of the absence of clouds. Note that the date choice is arbitrary. **(d)** and **(e)** are the same as **(a)** and **(b)** but for JRA-55. **(f)** satellite imagery retrieved from NASA's Global Imagery Browse Services for 2015-03-01 (daily composite) based on the MODIS false color 'snow RGB' (Bands 3-6-7). Note that the image is available only in regions experiencing direct sunlight on the day.

reliable and, at the same time, the reanalysis bias is larger. For this reason, identifying the occurrence of CSE in atmospheric reanalysis is a key step for an appropriate development and application of our correction strategy. In the framework of this study, two alternative approaches have been considered for this classification. The first identification approach is based on the total cloud cover (TCC) from atmospheric reanalyses. The TCC variable is defined as the proportion of a grid-cell covered by clouds, resulting in a single level field based on the clouds occurring at different vertical model levels by making assumptions on the degree of overlap/randomness between clouds at different heights. The performance of TCC for diagnosing CSE over the Arctic sea ice appears to be poor for the ERA5 reanalysis, which tends to overestimate the winter cloud cover (Gryning et al. 2020), but good for the JRA-55 product. This is shown in the qualitative comparison between the reanalyses TCC (Fig. 1; a and d), the number of measurements collected daily by the AVHRR sensor (Fig. 1c), and the satellite image retrieved by the MODIS instrument (Fig. 1f). Two more snapshots of the same panel are included in the supplementary materials (Figs. S1 and S2) to show that this condition is not only found in this specific case. Note that we do not use the number of measurements collected by the AVHRR sensors as the base for our cloud classification procedure because a low number of measurements can indicate a cloudy atmospheric state, but also an observational gap that has nothing to do with the cloud conditions. In contrast, the second classification approach relies on information about the atmospheric thermal (longwave) state, a variable typically described in atmospheric reanalyses both for a realistic atmosphere with clouds and for a hypothetical atmosphere without clouds. The difference between the all-sky and clear-sky surface downward thermal radiation (Δ_{STRD}) provides good indications of the presence of clouds for ERA5, as qualitatively illustrated by its good agreement with the observation density and the observed cloud state (Fig. 1; b, e, and f). Note that, due to the rapid evolution of the cloud as well as temperature states, analyzing snapshots from reanalysis and observations instead of long-term averages is more insightful for diagnosing similarities between weather patterns, an approach that we follow in the remainder of this manuscript.

After some manual calibration to identify the threshold values for each classification method, we decided to apply the temperature correction for the ERA5 reanalyses (i.e. assert a cloud free part) only to regions where $\Delta_{\mathrm{STRD}} \leq 15\,\mathrm{W/m^2}$. To avoid the development of nonphysical discontinuities in the surface temperature fields, we assign a temperature that proportionally combines corrected and original temperatures to transition regions where $15\,\mathrm{W/m^2} < \Delta_{\mathrm{STRD}} \leq 40\,\mathrm{W/m^2}$, building a transition zone between the corrected and uncorrected part of the domain. Finally, cloudy regions where $\Delta_{\mathrm{STRD}} > 40\,\mathrm{W/m^2}$ retain their uncorrected temperature. Given the good correspondence between TCC, cloud observations, and observation count for JRA-55, the application domain for this reanalysis product is defined based on the TCC variable. The corrected temperature is assigned where TCC $\leq$ 15%, the transition regime occurs where 15% $<$ TCC $\leq$ 70%, and finally no correction is applied where TCC $>$ 70%. In addition, for both reanalyses we further limit the correction to the sea ice pack (where sea ice concentration is larger than 80%), and locations with a reanalysis surface temperature lower than $-5\,^oC$. For higher temperatures, the surface temperature discrepancy between model and observation tends to be generally small. Under these conditions, we typically observe a low conductive heat flux because of the low temperature gradient between atmosphere, ice, and ocean, making a correction less relevant, and furthermore, there are not enough observations to perform a robust training of the correction model because of prevailing cloudy conditions in warm months.

d. The Correction Model Skill Score

We adopt the Correction Model Skill Score (CMSS) as a metric to measure the skill of the correction model in reducing the bias against independent observations.

$$CMSS = 1 - \frac{\mid SKT_{Cor} - SKT_{Obs} \mid}{\mid SKT_{Org} - SKT_{Obs} \mid}, \tag{1}$$

where SKT_{Cor} is the corrected reanalysis skin temperature, SKT_{Org} is the original reanalysis skin temperature, and SKT_{Obs} is the skin temperature measured independently. This metric should be interpreted as follow:

- CMSS = 1 means that the correction model brings the reanalysis temperature to match the observations and fully corrects the bias.
- For 0 < CMSS < 1, the correction model reduces the bias.
- CMSS = 0 means that the correction model has a neutral impact on the bias. Note that because the CMSS is an absolute metric, this case could refer both to the application of a null correction, but also to the introduction of a bias of the opposite sign.
- CMSS < 0 means that the correction model degrades the reanalysis.

3. Results

a. Characterization of the Temperature Bias and its Correction

The role of the atmospheric and sea ice predictors in shaping the skin temperature correction has been investigated during the training phase of the ML correction model. The relationship between the ERA-5 and JRA-55 temperature bias and the predictors is visualized in Fig. 2 (plots a, b, e, f). Only 10^5 randomly selected points out of the approximately 10^7 composing the test datasets are shown here to allow clearer visualization of the bias features. As a reminder, the test dataset is built with reanalysis data and observations from the years 2000 to 2009 that fulfill the clear sky classification and, for this reason, the considerations on the bias nature can only refer to the clear sky state, an essential condition for ensuring precise observations of the surface temperature. The temperature bias is defined as the difference between the reanalysis state and the observed temperature. As such, in the context of this study, a positive temperature bias indicates that the reanalysis product is warmer than the observations, while the opposite is true for a negative bias.

The emerging structure of the bias confirms the finding of previous studies and our physical understanding of the coupled atmospheric-sea ice system. The main features of the temperature bias are summarized in the following points:

- Large positive temperature biases are evident for cold reanalysis temperatures and low downward longwave radiation values, particularly for ERA5 (Fig. 2 a and e).
- Large positive temperature biases occur in regions with thick sea ice, thick snow, or a combination of both conditions (Fig. 2 b and f).
- Moderate negative biases tend to occur for thin sea ice, thin snow, or a combination of both conditions (Fig. 2 b and f).
- Despite the well recognizable features described in the previous points, the bias also shows a certain random error component that can be linked to inevitable differences between the observed and reanalysis state.

The mismatch between reanalysis and observations ranges approximately between -8^oC and $+2^oC$ for ERA5, and -8^oC and $+6^oC$ for JRA-55. These large values are in agreement with the estimates of previous studies. A comparison between ERA5 and JRA-55 reveals some differences in the relationship between the bias and the atmospheric predictors (Fig. 2 a and e). While the largest positive temperature bias in ERA5 is observed for cold temperatures (-40^oC to -25^oC), the situation is less obvious for JRA-55, which also exhibits a higher level of noise. Note that the truncation for temperature values above -5^oC (plots a, c, e, and g) is obtained by construction, as no correction is applied for temperatures warmer than -5^oC. For a given temperature, the spread

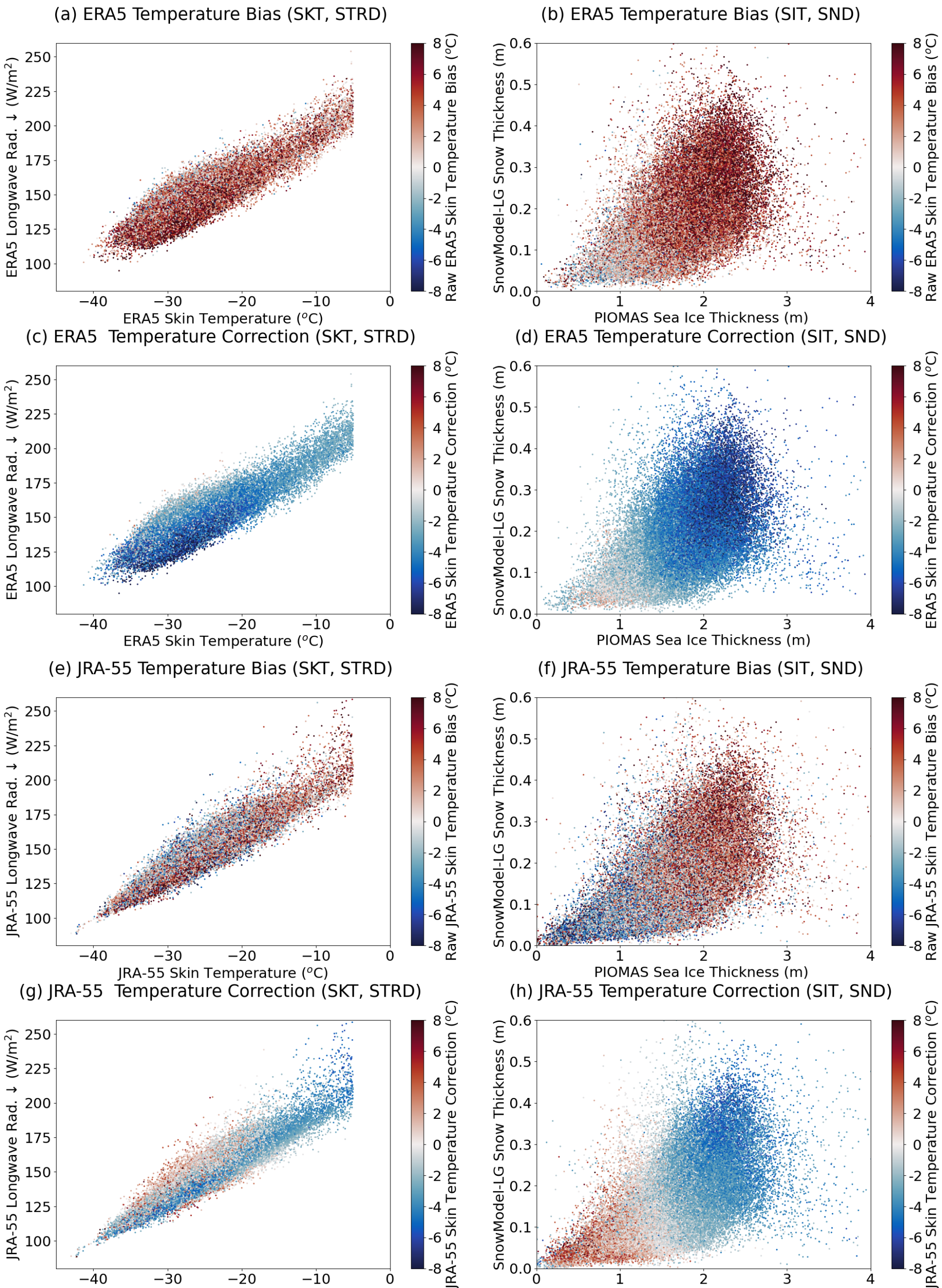

FIG. 2. Comparison between the skin temperature bias (reanalysis temperature minus observed temperature; **(a)**, **(b)**, **(e)**, **(f)**) and modelled skin temperature correction (output of the ML correction model; **(c)**, **(d)**, **(g)** to **(h)**). These color coded quantities are plotted as function of the atmospheric predictors SKT and STRD and the ice predictors SIT and SND.

of the downward longwave radiation values is bigger in ERA5 than in JRA-55 (y-axis in Fig. 2 a and e). When considering the sea ice predictors, the bias shows a functional relation to the sea ice thickness in both reanalyses, while the dependence on the snow depth is less pronounced and seems relevant only for sea ice thinner than 1 m. This is consistent with our physical understanding of the system: for thick sea ice, the effect of snow on heat conduction is small because the sea ice already saturates the insulation, while for thin sea ice the snow drives the conduction properties of the system.

The temperature correction predicted by the ML correction model is shown in Fig. 2 as a function of the four predictors (plots c, d, g, and h). Note that the same test points are displayed for the bias plots (first and third row) and correction plots (second and fourth row). Overall, the structure of the correction captures well the features of the original bias discussed in the previous paragraphs. The opposite sign of correction and bias makes physical sense and, ideally, a perfect correction would exactly cancel out the reanalysis bias. The predicted correction tends to be smooth and does not exhibit the same noise as the bias. On one hand, this is a positive feature and it indicates that the NN captures the systematic error while neglecting the random component. On the other hand, due to this behavior, the NN seems unable to correct extreme cases when the absolute difference between reanalysis and observed temperature is high. The latter is a feature of the correction model and not of the training procedure (i.e. it is not linked to size limitation in the training dataset or to the frequency of occurrence of these extreme events).

As the next step, we want to understand whether the correction learned by the ML model during the training phase can be applied to the reanalysis temperature field in a more operational setup, thus investigating if the corrected temperature fields retain the spatial coherency of the original reanalysis products, ideally also outside the training time window.

Maps a and d in Fig. 3 exhibit the original skin temperature field for ERA5 and JRA-55 respectively. Part of this discrepancy is simply explained by the different spatiotemporal resolutions of the two reanalyses (lower in JRA-55 than in ERA5). Nevertheless, another part originates from the different model physics and, in particular, for the resulting cloud states, with ERA5 featuring more clouds than JRA-55 (Fig. 1). Note that considering the same reanalysis snapshot in Figs. 1 and 3 allows us to relate the surface skin temperature and its correction to the cloud and downward longwave radiation state. While both maps show similar spatial features, they also reveal different

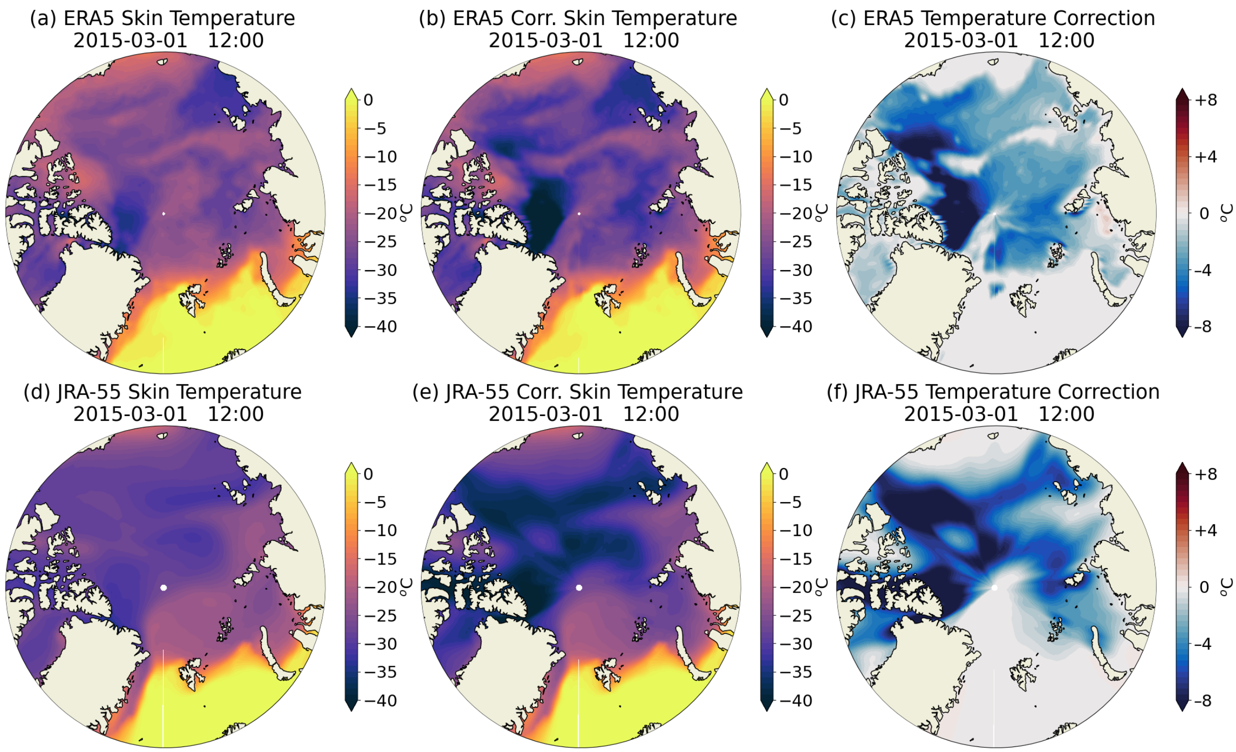

FIG. 3. **(a)** 2015-03-01 original ERA5 skin temperature over sea ice and open ocean. **(b)** 2015-03-01 corrected ERA5 skin temperature over sea ice and open ocean. **(c)** 2015-03-01 ERA5 temperature correction over sea ice. **(d)**, **(e)**, and **(f)** are respectively the same as **(a)**, **(b)**, and **(c)** but for the JRA-55 reanalysis.

temperatures. The warm regions ($-20^o C <$ SKT $< -15^o C$) are larger in ERA5 but, at the same time, the cold regions are also slightly colder for this dataset. The correction application leads to a marked cooling in the clear-sky portion of the domain. Note that the difference in the active correction domain for the two reanalyses, as well as the magnitude of the correction, is in part due to differences in the cloud state representation, in part to the application of different classification strategies for the clear sky state in reanalyses (Sec. 2c), and in part to the application of two different correction models. The locations on which the temperature correction is applied are generally continuous over relatively wide portions of the Arctic and evolve dynamically following the movement of large-scale weather systems. The presence of localized cloud formations and clear-sky gaps introduce heterogeneity to the active correction domain. This feature is particularly evident for ERA5, which can resolve smaller cloud formations due to the higher spatiotemporal resolution. No further unexpected spatial noise or sharp gradients emerges from the correction, indicating that the choices made concerning the application mask are reasonable. Overall, each reanalysis maintains consistency with its atmospheric state after the correction application.

b. Comparing the Corrected Skin Temperature to Independent In-situ Observations

A rigorous evaluation of the correction model skill mandates comparing the corrected temperatures with independent measurements, possibly outside the training decade. The meteorological dataset collected during the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) expedition in the winter of 2019–2020 (Shupe et al. 2022; Reynolds and Riihimaki 2019) provides an ideal basis for building this assessment. During MOSAiC, a set of longwave broadband up- and down-welling observations were made from a location on the sea ice. The surface skin temperature was derived from these measurements assuming a fixed surface emissivity of 0.985, which is reasonable for the winter observations used here.

As expected, Fig. 4a and b reveal large positive skin temperature biases for both the reanalyses when compared to the in-situ observations, particularly in association with clear sky conditions. The correction model performs reasonably well and tends to substantially mitigate the bias for ERA5, with a 27% average bias reduction, while the improvement is modest for JRA-55, with a 7% average bias reduction. The above reduction percentages have been quantified by computing the Mean Absolute Error (MAE) based on all the winter MOSAiC observations available from October 2019 to June 2020 (Tab. 1, columns 2 and 3 – *All Observations*), including instances of cloudy conditions when the temperature correction does not act. The error reduction for ERA5 and JRA-55 increases respectively to 32% and 10% when restricting the analysis only to clear-sky conditions according to each reanalysis classification (Tab. 1, columns 4 and 5 – *Clear-sky Observations*). The Pearson correlation between the reanalysis and observation time series is 0.89 for ERA5 and 0.75 for JRA-55, with negligible differences between the corrected and original cases. The complete MOSAiC temperature time series for ERA5 and JRA-55 are available in the supplementary materials (Fig. S3), while Fig. 4 focuses on four winter months only for better readability of the panel.

Comparing gridded reanalysis fields at relatively low resolution with single-point measurements is challenging and requires additional care to draw the correct conclusions. Firstly, reanalyses data represent spatially an average sea ice and snow state, while in-situ observations capture a unique

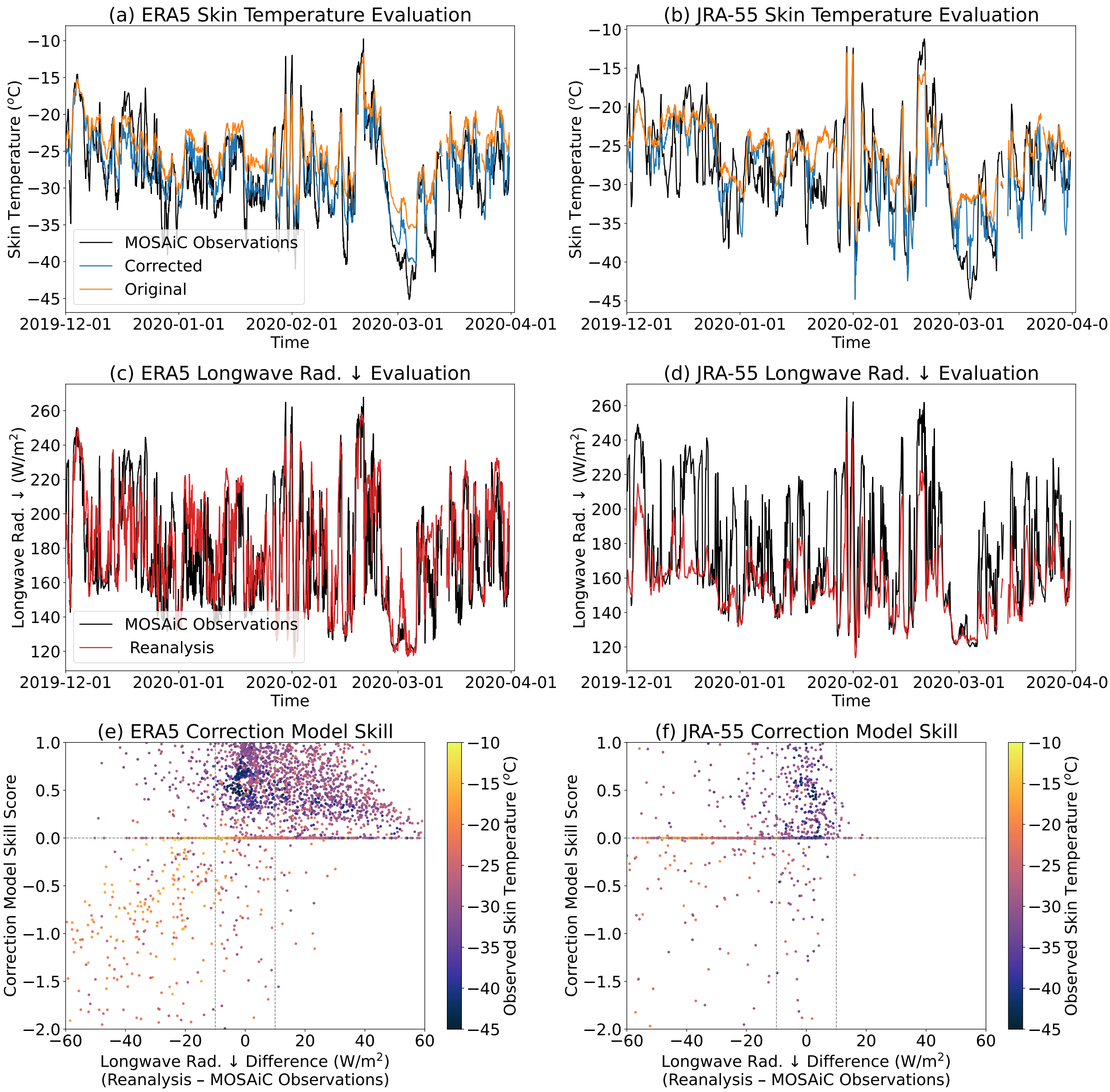

FIG. 4. **(a)** and **(b)**: Skin temperature measured during the MOSAiC expedition and estimates from the corrected and original reanalyses from 01-12-2019 to 31-03-2020. **(c)** and **(d)**: Same as **(a)** and **(b)**, but exhibiting the downward longwave radiation. **(e)** and **(f)**: Correction model skill score as function of the downward longwave radiation difference between reanalyses and MOSAiC observations. Note that the different point density in the two plots is due to the different time resolution of the reanalyses.

ice state. There is no straightforward way to accurately downscale the gridded data and account for this uncertainty. Secondly, the cloud state of in-situ observations and reanalysis should be

	All Observations		Clear-sky Observations		Compatible Observations	
	ERA5	**JRA-55**	**ERA5**	**JRA-55**	**ERA5**	**JRA-55**
Original	3.75 ^oC	3.52 ^oC	4.06 ^oC	3.83 ^oC	3.56 ^oC	4.41 ^oC
Corrected	2.75 ^oC	3.29 ^oC	2.75 ^oC	3.45 ^oC	1.80 ^oC	3.52 ^oC
Error Reduction	27%	7%	32%	10%	49%	20%

TABLE 1. Average temperatures mismatch between reanalysis and MOSAiC observations (October 2019 to June 2020) quantified by the Mean Absolute Error (MAE) metric for the corrected and original case considering all the available MOSAiC observations (columns 2 and 3), only clear-sky observations according to each reanalysis classification (columns 4 and 5), and only the observations with a longwave radiation state compatible with the reanalysis (columns 6 and 7).

similar for a meaningful comparison, which is not necessarily the case in our situation, as shown in Fig. 4c and d. Specifically, the STRD in JRA-55 is substantially lower than in the measurements when clouds are present (i.e. for the highest values in STRD), and also the ERA5 evaluation reveals differences in multiple instances. Therefore, we display the CMSS (Fig. 4; plots e and f) as a function of the downward longwave radiation difference between the two reanalyses and the MOSAiC observations (Δ_{STRD*}). We argue that the model skill is meaningful only when this difference is small (-10 W/m^2 < Δ_{STRD*} < 10 W/m^2). Under these conditions, the model skill scores are generally positive, with 49% bias reduction for ERA5 and 20% for JRA-55 (Tab. 1, columns 6 and 7 – *Compatible Observations*), and we observe only a few instances when the correction degrades the reanalysis. Outside this range, the skill score can capture a bias reduction or degradation for the wrong reasons.

Given the results that emerge from this independent evaluation, we believe that our method provides a useful correction for ERA5. However, for JRA-55, the correction performance is quite small. We expand on possible reasons for this discrepancy between the different reanalysis products below and discuss possible steps forward.

c. Spatiotemporal Variability of the Temperature Correction

Because of the rapid changes that the Arctic experienced during the last few decades, such as the decline of the sea ice extent and volume in response to the warming of both the near-surface atmosphere and the ocean, there are good reasons to believe that also the reanalysis skin temperature bias, as well as its correction, will present some trends and a certain level of spatiotemporal variability. This hypothesis is reasonable also given our understanding of the mechanism inducing the bias, which is ice thickness and temperature-dependent. For instance, the constant sea ice thickness assumption (e.g. 1.5m in ERA5) made in the reanalysis models, appears to be more compatible with the recent (post 2007) winter sea ice condition compared to those observed at the end of the 20$^{\text{th}}$ century. Similarly, for a given year and depending on the season, this assumption might be appropriate for certain Arctic locations while penalizing for others. We will begin exploring these aspects by making some consideration on the average spatial distribution of the correction during the different seasons.

Fig. 5 exhibits the 1981 to 2020 average temperature correction for the months December-January-February (DJF), March-April-May (MAM), and September-October-November (SON). Note that cloudy regions and open water regions, where the correction is zero, are also included in this spatiotemporal average. For both reanalyses, the correction exhibits a moderate seasonality. Specifically, it reaches a maximum in winter (DJF; Fig. 5 a and d), when the Arctic is colder and drier, and a minimum in the summer months, when by design no correction is applied because of too warm temperatures (maps not shown for June, July, and August). Furthermore, the fall correction (SON; Fig. 5 c and f) is smaller than the late winter/early spring one (MAM; Fig. 5 b and e), a fact that can be counter-intuitive given Arctic temperature similarities during these two periods, but that it is explained by the presence of thicker and thus more insulating snow and ice layers in MAM, which is conducive to the warm bias (see Fig. 2). Furthermore, given that zero correction regions are included in the average, this behavior can also be caused by different cloud and open water conditions in SON than in MAM, particularly for the most recent years. Both reanalyses feature a large negative correction over thick sea ice regions (north of the Canadian Archipelago and Greenland), and a smaller one (in absolute terms) in peripheral seas with a seasonal ice cover. A similar structure, including the differences between JRA-55 and ERA5, has been evidenced in the temperature bias quantification by Batrak and Müller (2019) (Fig. 3 of their paper; maps c and d), even though the comparison is possible only in qualitative terms due to the different periods and methodologies of our analyses. Even though instances of a positive correction up to $2^{o}C$ occur in single snapshots, particularly during the fall months in peripheral Arctic seas, these disappear in the multi-year, multi-month average of Fig. 5. A positive temperature correction instance can be observed in Fig. 3c along the Kara Sea coast, and it is linked to a sea ice divergence area which leads to a thinner sea ice and snow cover. Note that the overall corrections to ERA5 are slightly smaller than corrections to JRA-55, which might lead the reader to conclude that the original ERA5 temperature is closer to observed than JRA-55. However, this is not the case for the MOSAiC analysis (Tab. 1, row 1, columns 1 to 4), and this feature might be also explained by the effect of a larger cloudiness in ERA5 compared to JRA-55, hence less opportunity to correct the temperature field under the clear sky state.

The plot in Fig. 6a shows the annual cycle of the difference between the uncorrected and corrected atmospheric surface temperature averaged over the region north of 70N. In this context, positive difference values correspond to a negative correction as defined in Figs. 2 and 5. The results have been grouped in four different periods, roughly representative of the last four decades, to reveal the possible interannual trends of the correction. The seasonal cycle of the temperature difference confirms previous evidence that the correction reaches a maximum in winter and a minimum in the summer. Furthermore, a declining trend characterizes both the ERA5 (solid lines) and JRA-55 (dashed lines) corrections for the last decade (2010—2019; red lines). During the last decade (2010–2019), the average correction for both reanalyses becomes almost zero for the transitions months of May and October, demonstrating a generalized time reduction of the active correction season as the sea ice thickness decreases and the Arctic warms. During the winter months (February to April), the multi-decadal evolution of the reanalysis correction before 2010 becomes less obvious, likely due to a strong reduction of the heat conduction through the ice after a certain effective conductivity threshold (defined by the sea ice and snow thickness) is reached.

Applying the correction to the reanalyses fields tends on average to cool the climatological temperature state over the Arctic sea ice, and this could in principle impact the reanalysis representation of the warming that the Arctic experienced during the last decades. We investigate this aspect in Fig. 6 (plots b and c), where the anomalies for the corrected and uncorrected skin temperatures (computed against their climatological reference based on the period 1981-2010) are respectively displayed for the ERA5 (plot b) and JRA-55 (plot c) reanalyses. Note that each anomaly time series is built by subtracting its individual climatological state, and not a common one. For both reanalyses, the anomaly variability is similar for the original (red lines) and the corrected data (blue lines), with only small differences between the two. The warming trend of the original product is slightly smaller than that of the corrected product for both reanalyses: ERA5 exhibits a warming of $0.98\frac{K}{10y}$ for the corrected case and $0.82\frac{K}{10y}$ for the uncorrected case. JRA-55 exhibits a warming of $0.92\frac{K}{10y}$ for the corrected case and $0.80\frac{K}{10y}$ for the uncorrected case. Thus, the correction impact on the warming trend for JRA-55 75% of that of ERA5. This difference is still relatively small ($\sim$ 10% to 20%) if compared to the absolute magnitude of the warming signal and in line with the trend of differences between the two reanalysis products.

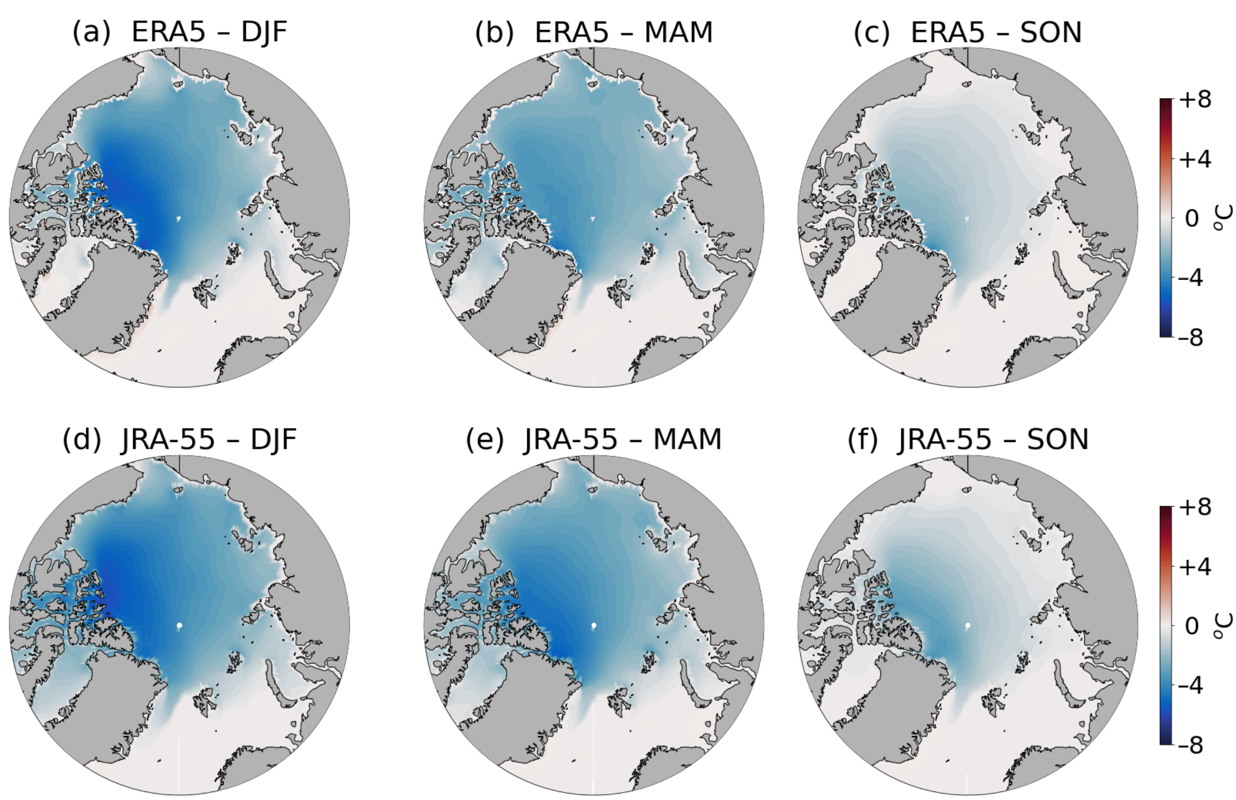

FIG. 5. 1981 to 2018 average temperature correction for the months December-January-February (DJF), March-April-May (MAM), and September-October-November (SON). The ERA5 and JRA-55 maps are respectively grouped in the upper and bottom row. The summer months are not shown because the correction is zero. All the maps share the same color scheme illustrated by the color bars on the right. Note that, in agreement with Fig. 2, the sign of the correction is opposite of that of the bias.

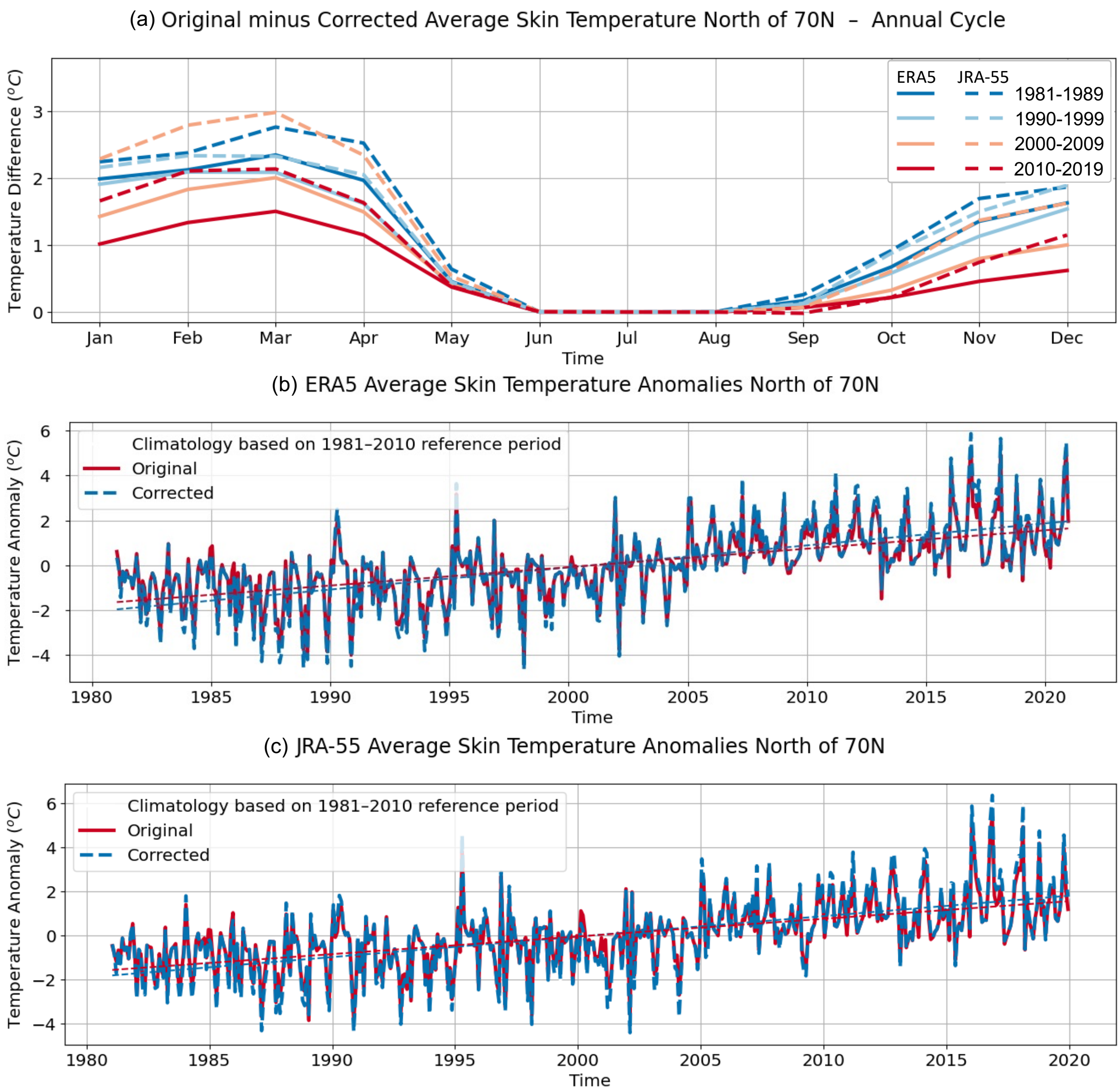

FIG. 6. **(a)** Annual cycle averaged over four decades of the difference between the original (uncorrected) and the corrected ERA5 (solid lines) and JRA-55 (dashed lines) skin temperatures averaged over the regions north of 70N. **(b)** and **(c)** Corrected (blue dashed lines) and original (red lines) ERA5 and JRA-55 skin temperature anomalies computed against their own climatological reference based on the period 1981-2010. The dashed straight lines quantify the average warming trend experienced by the Arctic over the period under consideration.

4. Discussion

a. Limitations of the Proposed Bias Correction Strategy

The bias correction strategy presented in this study proved to be effective in partially correcting the near-surface temperature bias that affects the current generation of atmospheric reanalysis in the Arctic region. Nevertheless, some limitations associated with our methodology deserve some more in-depth discussion.

The first caveat of our approach is that the ML correction model is trained on a limited portion of the reanalysis period (2000 to 2009) while being applied also to previous or future decades experiencing different conditions (i.e. on average colder temperatures and thicker sea ice and snow before 2000, and the opposite after 2010). We argue that this assumption is acceptable, given that our correction model design relies on state-dependent predictors and not on spatiotemporal information such as the location and the time of the year—also legitimate predictors that would however strongly bind the model to the background climate state. Furthermore, the misrepresentation of the conductive heat flux through sea ice and snow, which is the mechanism at the heart of the observed bias, tends to saturate for thick ice and snow, for which the conductive heat flux becomes very small. Nevertheless, we cannot exclude that the correction is sub-optimal for sea ice regimes underrepresented in the training dataset, such as very thick ice conditions, and we can only rely on the extrapolation capabilities of the ML model under these conditions. Encouraging indications of the robustness of our approach to this kind of issue come from the self-emerging declining trend of the correction for both the reanalyses products considered, which highlight the dependence of the model on the sea ice state, and the convincing comparison to MOSAiC in-situ observations outside of the training window.

A second point worth discussing is the fact that the correction model relies entirely on reanalysis products, which have themselves well-known shortcomings. For example, in terms of the ice predictors, the limitations of the PIOMAS product, which consistently underestimates the sea ice thickness in regions of thicker ice and overestimates it in regions of thinner ice, are well documented in the literature (Labe et al. 2018). The physical sophistication of the SnowModel-LG thickness product is remarkable, but this product is by design impacted by errors in the snow precipitation and sea ice drift description used to force the reanalysis model. While alternative direct Arctic-wide observations of the snow thickness are presently not available, remote sensing sea ice thickness observations (e.g. from EnviSat, CryoSat-2, SMOS, and IceSat2 satellites) and reanalyses (Mu et al. 2020, 2022) have become available for the past 20 years. While we considered employing some of these products as an alternative to PIOMAS, we decided against this approach in order to apply the correction model consistently over the entire reanalysis period with no spatiotemporal gaps due to missing observations. A complementary correction approach considered for this study consisted of nudging the reanalysis surface state to the satellite observations when these were available. Even though this would have certainly led to good temperature estimates in areas with a high density of observations, and also limited the episodes of bias degradation associated with the application of the correction model, we decided against this strategy to avoid the introduction of inconsistencies in the corrected reanalysis field, as observations are not regularly available over the whole domain, and they are temporally incompatible with the reanalysis products (daily versus sub-daily representation).

The discussed bias correction approach targets the Arctic, while we expect similar biases to emerge also for the Antarctic sea ice. The main motivation for this is the absence of ice predictors; with no reliable long term Antarctic sea ice and snow thickness estimates our correction model would lose a substantial portion of its skill, a fact that prevents us from even testing our Arctic trained correction on the Antarctic domain. Furthermore, the compatibility of the reanalyses with the true atmospheric state is strongly linked to the number of observations assimilated in the forecast system. A better reanalysis quality for more recent years than the past should thus be expected due to the advances in observational techniques. While under clear-sky conditions the Arctic boundary layer is strongly decoupled from the rest of the atmosphere and poorly characterized by observations also for recent years, the locations at which clear-sky conditions occur can be affected by the quality of the circulation in the reanalysis. Correcting for circulation issues in reanalyses goes beyond the scope of this study, and this aspect should be kept in mind when using these products in polar regions, with or without bias correction.

A further aspect to consider is the difference between skin temperature and 2m temperature in reanalysis products. Given that the observed temperatures used to quantify the reanalysis bias are representative of the surface layer, the resulting correction is also applied to the skin temperature of the reanalysis. However, most of the reanalysis temperature applications in polar regions are based on the 2m temperature, including the forcing fields for sea ice and ocean models. To maintain consistency between the reanalysis fields, we transfer the skin temperature correction to the 2m temperature variable by assuming that the temperature difference between these two model levels would remain unchanged. The robustness of this assumption is hard to prove, given that the stratification of the near-surface atmosphere cannot be observed from remote sensing products, and thus its characterization mostly relies on local measurements. Other reanalysis variables defining the surface energy budget, such as the surface turbulent heat flux and the upwelling longwave radiation, must also be affected by biases because the uncorrected skin temperature is biased. Both these quantities have an impact on boundary layer and cloud processes. Once the skin temperature is corrected using the method presented here, it is then inconsistent with the other uncorrected terms in the reanalyses surface energy balance, and this aspect should be considered carefully to avoid misuse of the corrected product.

The correction application domain is tightly linked to the cloud state, and the assumptions made in the classification of clear-sky versus cloudy regions impact the correction. Unfortunately, the lack of direct surface observations in cloudy conditions made an extension of the ML model to the cloudy state impossible. Also, in these conditions there are many more physical processes involved, (e.g. cloud radiative properties) which would make the ML model training more challenging. In the attempt to overcome this limitation, during the preliminary phase of our work, we tried to integrate the remote sensing observations with arguably more precise in-situ measurements collected by automatic buoys and weather stations deployed on the Arctic sea ice. These observations are less abundant than satellite products, but provide a more complete overview of the surface temperature state in the Arctic, also covering earlier decades, cloudy conditions, as well as being available for the Southern Ocean sea ice. However, comparing localized observations representative of a very specific sea ice state to gridded products that capture an average sea ice state representative of an area spanning several kilometers, proved to be unfeasible, as we also argue in Sec. b.

Finally, the correction skill difference between ERA5 and JRA-55 deserves additional discussion. The model skill that emerges from the comparison to independent MOSAiC observations reveals better performances for ERA5 than JRA-55. We speculatively attribute the low JRA-55 skill to lower synoptic and moisture compatibility of this reanalysis with the true atmospheric state, as suggested by the lower temporal correlation with the MOSAiC observations and the downward longwave radiation analysis. First, the discrepancy impacts the correction at the model training stage, as the learned bias signal generates not only from the snow-related mechanism but also from unrelated sources. Second, the discrepancy results in penalization at the evaluation stage, as the correction can exacerbate the bias if observations and reanalysis are in different regimes. Nevertheless, further analyses are needed to quantitatively verify the previous statement and formulate a correct attribution of the correction skill difference.

b. Comparing the Bias Correction Methodology to Previous Correction Strategies

Even though a clear understanding of the physical mechanism responsible for the winter temperature bias in atmospheric reanalysis has been uncovered only in recent years, the existence of the bias itself has been established earlier and several measures have been taken for mitigating its effect. In particular, the ocean and sea ice modeling community realized that employing uncorrected reanalysis temperature fields as forcing (i.e. boundary conditions) for regional and global sea ice and ocean general circulation models leads to an unsatisfactory representation of the sea ice (mainly not enough sea ice formation during winter), with errors propagating also to other seasons and ultimately to the oceanic circulation in the Arctic and beyond. Two alternative approaches can be taken to mitigate this problem: 1. tuning underconstrained key model parameters to partially compensate the forcing effect (Zampieri et al. 2021; Sumata et al. 2019), for example by increasing the sea ice and snow conductivity to foster the heat conduction through the sea ice system, and 2. calibrating the reanalysis, and thus following the same reasoning that motivated this study. The latter approach has been attempted by the DRAKKAR project, which develops consistent global forcing datasets based on a combination of ECMWF reanalysis and observed flux data, called Drakkar Forcing Sets (DFS). To correct the ERA40 warm Arctic bias, the DFS adopts a full spatially dependent monthly rescaling of ERA40 air temperature over ice-covered regions north of 70°N, using a monthly climatological sea-ice mask (Brodeau et al. 2010), a stratagem that follows the work of Large and Yeager (2004) and Large and Yeager (2008) in the context of the Coordinated Ocean Reference Experiments and the "CORE2" forcing. More recently, the community participating in the Ocean Models Intercomparison Project (OMIP) proposed a calibration strategy for the JRA-55 temperature in the Arctic (Tsujino et al. 2018) based on data from the International Arctic Buoy Programme (IABP) / Polar Exchange at the Sea Surface (POLES) (IABP-NPOLES; (Rigor et al. 2000)), and implemented in the JRA-55-do forcing.

The previously mentioned strategies can be classified as climatological calibration, meaning that they aim to a correct climatological representation of the temperature in the Arctic. However, we argue that our correction approach, compared to the previous attempts, brings a higher level of sophistication for three main reasons:

1. The correction is state-dependent, meaning that it is coherent with the reanalyzed sea ice conditions and with the local weather. It favors clear-sky conditions, in agreement with the observation-based characterization of the reanalysis bias. Furthermore, its predictors can be associated with the physical mechanism causing the bias in the first place, which is the misrepresentation of the conductive heat flux through the snow and sea ice.

2. Even though the reanalysis bias in the Arctic is on average warm, our model is able to correct also less common occurrences of cold biases occurring on thin ice, mostly at the beginning of the freezing season.

3. A self-emerging property of the correction is its declining trend for the last decade, which is compatible with our physical understanding of the bias and with the changing sea ice conditions in the Arctic due to global warming.

In addition, a characteristic of our correction is that, similarly to the climatological calibration approaches, it has only a minor impact on the reanalysis representation of the near-surface warming trend of the Arctic observed in the past four decades. A quantitative comparison of our correction strategy with previous efforts falls outside the scope of this work.

5. Conclusion

In this study, we have presented a machine learning correction model that reduces the (mostly warm) winter bias over the Arctic sea ice in uncoupled atmospheric reanalyses due to a misrepresentation of the conductive heat flux through the sea ice and snow. Our work focused on the widely used ERA5 and JRA-55 products, but no constraint would prevent the model from being trained also on other reanalysis products, as well as on coupled forecast systems exhibiting similar biases. The correction relies on four reanalysis predictors, which have been chosen because they are skillful and linked to the physical mechanism that causes the bias. These are the reanalysis surface temperature itself, the downward longwave (or thermal) radiation reaching the surface, the sea ice thickness, and the snow thickness. The skill of the correction model is investigated by comparing the original and corrected reanalyses to independent in-situ measurements from the

MOSAiC campaign. This comparison revealed an overall positive impact of the correction, with a substantial reduction of the bias and only limited instances of degradation for ERA5, while the improvement is modest for JRA-55. The self-emerging properties of the correction are compatible with our understanding of the bias and of the ice system: the correction varies seasonally with a maximum in winter and a minimum in summer, it is spatially heterogeneous and on average stronger on thicker sea ice, and finally, it shows a declining trend linked to the sea ice reduction and warming of the Arctic. Overall, the ML correction results confirm the physical understanding of the bias.

We envisage that the correction presented in this study will find its main application in support of uncoupled sea ice and ocean simulations that rely on reanalysis fields as atmospheric boundary conditions. A better representation of the near-surface weather could be beneficial for a correct simulation of the Arctic sea ice and should reduce the use of nonphysical tuning choices aiming at compensating the reanalyses bias, rather than at an accurate simulation of the sea ice processes. In this context, more research is needed to understand the impact of the corrected fields on model simulations, and an in-depth evaluation of these aspects, as well as a quantitative comparison with previous reanalysis-based forcing fields, is out of the scope of this work.

Finally, we argue that the state-dependent approach to bias-correct reanalysis fields that was followed in this study is beneficial compared to simpler climatological calibration techniques, and we expect that similar correction models could be adapted also for other reanalysis variables affected by bias related to model deficiencies. The MOSAiC-based skill assessment presented in this study reveals that part of the bias remains despite our correction, and further efforts are needed, both in the context of coupled model development and post-processing, for improving the quality of atmospheric reanalysis over sea ice. For this reason, developing a correction that directly targets the mechanism generating the bias can be informative and guide future development efforts to improve the realism of the atmospheric reanalysis system, in the Arctic and beyond.

Acknowledgments. As part of the Virtual Earth System Research Institute (VESRI), funding for the Multiscale Machine Learning In coupled Earth System Modeling (M2LInES) project was provided to Lorenzo Zampieri and Marika Holland by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program. Observations used here from the MOSAiC 2019–2020 expedition were made by the Atmospheric Radiation Measurement (ARM) User Facility, a U.S. Department of Energy (DOE) Office of Science User Facility Managed by the Biological and Environmental Research Program. M.D.S. was supported by the DOE (DE_SC0021341), NSF (OPP-1724551), and NOAA (NA22OAR4320151). Finally, we are grateful for the comments and suggestions received from three anonymous reviewers, which greatly improved the manuscript.

Data availability statement. The reanalysis data and observations used in this study are all freely available. The ERA5 reanalysis data can be downloaded from the Copernicus Climate Change Service (C3S) Climate Date Store `https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview`. The JRA-55 reanalysis data can be downloaded from the NCAR/UCAR Research Data Archive `https://rda.ucar.edu/datasets/ds628.0/`. The OSI SAF sea ice concentration observations are available on the following pages: `https://osi-saf.eumetsat.int/products/osi-450` and `https://osi-saf.eumetsat.int/products/osi-430-b-complementing-osi-450`. The Arctic sea and sea ice surface temperature observations can be downloaded from Centre for Environmental Data Analysis (CEDA) archive `https://catalogue.ceda.ac.uk/uuid/60b820fa10804fca9c3f1ddfa5ef42a1?search_url=\%2F\%253Fq\%253DEUSTACE\%2BAVHRR\%26results_per_page\%253D20\%26sort_by\%253Drelevance`. The PIOMAS gridded sea ice concentration and volume per unit are can be downloaded from the Polar Science Center website `http://psc.apl.uw.edu/research/projects/arctic-sea-ice-volume-anomaly/data/`. The SnowModel-LG snow depth gridded data can be downloaded from the following NSIDC page `https://nsidc.org/data/nsidc-0758/versions/1`. The temperature and radiation observations from the MOSAiC campaign that have been employed in this study are based on Reynolds and Riihimaki (2019). The corrected reanalyses temperature products are stored at the Globally Accessible Data Environment (GLADE) managed by the National Center for Atmospheric Research and can be downloaded through Globus at https://app.globus.org/file-manager?origin_id=abf82ebb-21d6-4324-9d1a-59dc23332bee&origin_path=\%2F.

References

Arduini, G., S. Keeley, J. J. Day, I. Sandu, L. Zampieri, and G. Balsamo, 2022: On the importance of representing snow over sea-ice for simulating the arctic boundary layer. *Journal of Advances in Modeling Earth Systems*, **14 (7)**, https://doi.org/10.1029/2021ms002777, URL https://doi.org/10.1029/2021ms002777.

Batrak, Y., and M. Müller, 2019: On the warm bias in atmospheric reanalyses induced by the missing snow over arctic sea-ice. *Nature Communications*, **10 (1)**, https://doi.org/10.1038/s41467-019-11975-3, URL https://doi.org/10.1038/s41467-019-11975-3.

Brodeau, L., B. Barnier, A.-M. Treguier, T. Penduff, and S. Gulev, 2010: An ERA40-based atmospheric forcing for global ocean circulation models. *Ocean Modelling*, **31 (3-4)**, 88–104, https://doi.org/10.1016/j.ocemod.2009.10.005, URL https://doi.org/10.1016/j.ocemod.2009.10.005.

Copernicus Climate Change Service, 2021: Arctic regional reanalysis on single levels from 1991 to present. ECMWF, URL https://cds.climate.copernicus.eu/doi/10.24381/cds.713858f6, https://doi.org/10.24381/CDS.713858F6.

Day, J. J., S. Keeley, G. Arduini, L. Magnusson, K. Mogensen, M. Rodwell, I. Sandu, and S. Tietsche, 2022: Benefits and challenges of dynamic sea ice for weather forecasts. *Weather and Climate Dynamics*, **3 (3)**, 713–731, https://doi.org/10.5194/wcd-3-713-2022, URL https://doi.org/10.5194/wcd-3-713-2022.

Dybkjær, G., R. Tonboe, and J. L. Høyer, 2012: Arctic surface temperatures from metop AVHRR compared to in situ ocean and land data. *Ocean Science*, **8 (6)**, 959–970, https://doi.org/10.5194/os-8-959-2012, URL https://doi.org/10.5194/os-8-959-2012.

Gryning, S.-E., E. Batchvarova, R. Floors, C. Münkel, H. Skov, and L. L. Sørensen, 2020: Observed and modelled cloud cover up to 6 km height at station nord in the high arctic. *International Journal of Climatology*, **41 (3)**, 1584–1598, https://doi.org/10.1002/joc.6894, URL https://doi.org/10.1002/joc.6894.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146 (730)**, 1999–2049, https://doi.org/10.1002/qj.3803, URL https://doi.org/10.1002/qj.3803.

Høyer, J. L., P. L. Borgne, and S. Eastwood, 2014: A bias correction method for arctic satellite sea surface temperature observations. *Remote Sensing of Environment*, **146**, 201–213, https://doi.org/10.1016/j.rse.2013.04.020, URL https://doi.org/10.1016/j.rse.2013.04.020.

Høyer, J. L., and J. She, 2007: Optimal interpolation of sea surface temperature for the north sea and baltic sea. *Journal of Marine Systems*, **65 (1-4)**, 176–189, https://doi.org/10.1016/j.jmarsys.2005.03.008, URL https://doi.org/10.1016/j.jmarsys.2005.03.008.

Hoyer, S., and J. Hamman, 2017: xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software*, **5 (1)**, https://doi.org/10.5334/jors.148, URL https://doi.org/10.5334/jors.148.

Høyer, J. L., G. Dybkjær, S. Eastwood, and K. S. Madsen, 2019: Eustace/aasti: Global clear-sky ice surface temperature data from the avhrr series on the satellite swath with estimates of uncertainty components, v1.1, 2000-2009. Centre for Environmental Data Analysis (CEDA), URL https://catalogue.ceda.ac.uk/uuid/60b820fa10804fca9c3f1ddfa5ef42a1, https://doi.org/10.5285/60B820FA10804FCA9C3F1DDFA5EF42A1.

Jung, T., and Coauthors, 2016: Advancing polar prediction capabilities on daily to seasonal time scales. *Bulletin of the American Meteorological Society*, **97 (9)**, 1631–1647, https://doi.org/10.1175/bams-d-14-00246.1, URL https://doi.org/10.1175/bams-d-14-00246.1.

Keeley, S., and K. Mogensen, 2018: Dynamic sea ice in the ifs. *ECMWF Newsletter*, https://doi.org/10.21957/4SKA25FURB, URL https://www.ecmwf.int/node/18874.

Kobayashi, S., and Coauthors, 2015: The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan. Ser. II*, **93 (1)**, 5–48, https://doi.org/10.2151/jmsj.2015-001, URL https://doi.org/10.2151/jmsj.2015-001.

Labe, Z., G. Magnusdottir, and H. Stern, 2018: Variability of arctic sea ice thickness using PIOMAS and the CESM large ensemble. *Journal of Climate*, **31 (8)**, 3233–3247, https://doi.org/10.1175/jcli-d-17-0436.1, URL https://doi.org/10.1175/jcli-d-17-0436.1.

Large, W., and S. Yeager, 2004: Diurnal to decadal global forcing for ocean and sea-ice models: The data sets and flux climatologies. *NCAR Technical Notes*, https://doi.org/10.5065/D6KK98Q6, URL http://opensky.ucar.edu/islandora/object/technotes:434.

Large, W. G., and S. G. Yeager, 2008: The global climatology of an interannually varying air–sea flux data set. *Climate Dynamics*, **33 (2-3)**, 341–364, https://doi.org/10.1007/s00382-008-0441-3, URL https://doi.org/10.1007/s00382-008-0441-3.

Lindsay, R., M. Wensnahan, A. Schweiger, and J. Zhang, 2014: Evaluation of seven different atmospheric reanalysis products in the arctic. *Journal of Climate*, **27 (7)**, 2588–2606, https://doi.org/10.1175/jcli-d-13-00014.1, URL https://doi.org/10.1175/jcli-d-13-00014.1.

Liston, G. E., P. Itkin, J. Stroeve, M. Tschudi, J. S. Stewart, S. H. Pedersen, A. K. Reinking, and K. Elder, 2020: A lagrangian snow-evolution system for sea-ice applications (SnowModel-LG): Part I—model description. *Journal of Geophysical Research: Oceans*, **125 (10)**, https://doi.org/10.1029/2019jc015913, URL https://doi.org/10.1029/2019jc015913.

Liston, G. E., C. Polashenski, A. Rösel, P. Itkin, J. King, I. Merkouriadi, and J. Haapala, 2018: A distributed snow-evolution model for sea-ice applications (SnowModel). *Journal of Geophysical Research: Oceans*, **123 (5)**, 3786–3810, https://doi.org/10.1002/2017jc013706.

Mu, L., L. Nerger, J. Streffing, Q. Tang, B. Niraula, L. Zampieri, S. N. Loza, and H. F. Goessling, 2022: Sea-ice forecasts with an upgraded AWI coupled prediction system. *Journal of Advances in Modeling Earth Systems*, **14 (12)**, https://doi.org/10.1029/2022ms003176, URL https://doi.org/10.1029/2022ms003176.

Mu, L., and Coauthors, 2020: Toward a data assimilation system for seamless sea ice prediction based on the AWI climate model. *Journal of Advances in Modeling Earth Systems*, **12 (4)**, https://doi.org/10.1029/2019ms001937, URL https://doi.org/10.1029/2019ms001937.

Nielsen-Englyst, P., J. L. Høyer, K. S. Madsen, R. T. Tonboe, G. Dybkjær, and S. Skarpalezos, 2021: Deriving Arctic 2 m air temperatures over snow and ice from satellite surface temperature measurements. *The Cryosphere*, **15 (7)**, 3035–3057, https://doi.org/10.5194/tc-15-3035-2021, URL https://doi.org/10.5194/tc-15-3035-2021.

Onogi, K., and Coauthors, 2007: The JRA-25 reanalysis. *Journal of the Meteorological Society of Japan. Ser. II*, **85 (3)**, 369–432, https://doi.org/10.2151/jmsj.85.369, URL https://doi.org/10.2151/jmsj.85.369.

Paszke, A., and Coauthors, 2019: Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 8024–8035, URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Rasmussen, T. A. S., J. L. Høyer, D. Ghent, C. E. Bulgin, G. Dybkjær, M. H. Ribergaard, P. Nielsen-Englyst, and K. S. Madsen, 2018: Impact of assimilation of sea-ice surface temperatures on a coupled ocean and sea-ice model. *Journal of Geophysical Research: Oceans*, **123 (4)**, 2440–2460, https://doi.org/10.1002/2017jc013481, URL https://doi.org/10.1002/2017jc013481.

Reynolds, R., and L. Riihimaki, 2019: Arm: Icerad. Atmospheric Radiation Measurement (ARM) Archive, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (US); ARM Data Center, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), URL https://www.osti.gov/servlets/purl/1814821/, https://doi.org/10.5439/1814821.

Rigor, I. G., R. L. Colony, and S. Martin, 2000: Variations in surface air temperature observations in the arctic, 1979–97. *Journal of Climate*, **13 (5)**, 896–914, https://doi.org/10.1175/1520-0442(2000)013⟨0896:visato⟩2.0.co;2, URL https://doi.org/10.1175/1520-0442(2000)013⟨0896:visato⟩2.0.co;2.

Serreze, M. C., A. P. Barrett, A. G. Slater, M. Steele, J. Zhang, and K. E. Trenberth, 2007: The large-scale energy budget of the arctic. *Journal of Geophysical Research*, **112 (D11)**, https://doi.org/10.1029/2006jd008230, URL https://doi.org/10.1029/2006jd008230.

Shupe, M. D., and Coauthors, 2022: Overview of the MOSAiC expedition: Atmosphere. *Elementa: Science of the Anthropocene*, **10 (1)**, https://doi.org/10.1525/elementa.2021.00060, URL https://doi.org/10.1525/elementa.2021.00060.

Sumata, H., F. Kauker, M. Karcher, and R. Gerdes, 2019: Simultaneous parameter optimization of an arctic sea ice–ocean model by a genetic algorithm. *Monthly Weather Review*, **147 (6)**, 1899–

1926, https://doi.org/10.1175/mwr-d-18-0360.1, URL https://doi.org/10.1175/mwr-d-18-0360.1.

Thielke, L., M. Huntemann, S. Hendricks, A. Jutila, R. Ricker, and G. Spreen, 2022: Sea ice surface temperatures from helicopter-borne thermal infrared imaging during the MOSAiC expedition. *Scientific Data*, **9 (1)**, https://doi.org/10.1038/s41597-022-01461-9, URL https://doi.org/10.1038/s41597-022-01461-9.

Tjernström, M., and R. G. Graversen, 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, https://doi.org/10.1002/qj.380, URL https://doi.org/10.1002/qj.380.

Tsujino, H., and Coauthors, 2018: JRA-55 based surface dataset for driving ocean–sea-ice models (JRA55-do). *Ocean Modelling*, **130**, 79–139, https://doi.org/10.1016/j.ocemod.2018.07.002, URL https://doi.org/10.1016/j.ocemod.2018.07.002.

Zampieri, L., H. F. Goessling, and T. Jung, 2018: Bright prospects for arctic sea ice prediction on subseasonal time scales. *Geophysical Research Letters*, **45 (18)**, 9731–9738, https://doi.org/10.1029/2018gl079394, URL https://doi.org/10.1029/2018gl079394.

Zampieri, L., H. F. Goessling, and T. Jung, 2019: Predictability of antarctic sea ice edge on subseasonal time scales. *Geophysical Research Letters*, **46 (16)**, 9719–9727, https://doi.org/10.1029/2019gl084096, URL https://doi.org/10.1029/2019gl084096.

Zampieri, L., F. Kauker, J. Fröhle, H. Sumata, E. C. Hunke, and H. F. Goessling, 2021: Impact of sea-ice model complexity on the performance of an unstructured-mesh sea-ice/ocean model under different atmospheric forcings. *Journal of Advances in Modeling Earth Systems*, **13 (5)**, https://doi.org/10.1029/2020ms002438, URL https://doi.org/10.1029/2020ms002438.

Zhang, J., and D. A. Rothrock, 2003: Modeling global sea ice with a thickness and enthalpy distribution model in generalized curvilinear coordinates. *Monthly Weather Review*, **131 (5)**, 845–861, https://doi.org/10.1175/1520-0493(2003)131⟨0845:mgsiwa⟩2.0.co;2, URL https://doi.org/10.1175/1520-0493(2003)131⟨0845:mgsiwa⟩2.0.co;2.