Surface turbulent fluxes from the MOSAiC campaign predicted by machine learning

Donald P. Cummins$^1$, Virginie Guemas$^2$, Christopher J. Cox$^3$, Michael R Gallagher$^4$, and Matthew D. Shupe$^5$

$^1$CNRM, Université de Toulouse, Météo-France, CNRS
$^2$Barcelona Supercomputing Center
$^3$NOAA-PSL
$^4$CIRES/NOAA-ESRL
$^5$University of Colorado Boulder

August 3, 2023

Abstract

Reliable boundary-layer turbulence parametrizations for polar conditions are needed to reduce uncertainty in projections of Arctic sea ice melting rate and its potential global repercussions. Surface turbulent fluxes of sensible and latent heat are typically represented in climate models using bulk formulae based on the Monin-Obukhov Similarity Theory (MOST), sometimes finely tuned to high stability conditions and the potential presence of sea ice. In this study, we test the performance of new, machine-learning (ML) flux parametrizations, using an advanced polar-specific bulk algorithm as a baseline. Neural networks, trained on observations from previous Arctic campaigns, are used to predict surface turbulent fluxes measured over sea ice as part of the recent MOSAiC expedition. The ML parametrizations outperform the bulk at the MOSAiC sites, with RMSE reductions of up to 70 percent. We provide a plug-in Fortran implementation of the neural networks for use in climate models.
Surface turbulent fluxes from the MOSAiC campaign
predicted by machine learning

Donald P. Cummins¹, Virginie Guemas¹, Christopher J. Cox², Michael R. Gallagher²,³, Matthew D. Shupe²,³

¹CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France
²National Oceanic and Atmospheric Administration, Physical Sciences Laboratory, Boulder, CO, USA
³Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA

Key Points:

• Neural networks trained on previous Arctic campaigns predict surface turbulent fluxes from MOSAiC more accurately than bulk methods.
• Updated parametrizations using the MOSAiC data have been developed and implemented in Fortran for deployment in climate models.

Corresponding author: Donald P. Cummins, donald.cummins@meteo.fr
Abstract

Reliable boundary-layer turbulence parametrizations for polar conditions are needed to reduce uncertainty in projections of Arctic sea ice melting rate and its potential global repercussions. Surface turbulent fluxes of sensible and latent heat are typically represented in climate models using bulk formulae based on the Monin-Obukhov Similarity Theory (MOST), sometimes finely tuned to high stability conditions and the potential presence of sea ice. In this study, we test the performance of new, machine-learning (ML) flux parametrizations, using an advanced polar-specific bulk algorithm as a baseline. Neural networks, trained on observations from previous Arctic campaigns, are used to predict surface turbulent fluxes measured over sea ice as part of the recent MOSAiC expedition. The ML parametrizations outperform the bulk at the MOSAiC sites, with RMSE reductions of up to 70 percent. We provide a plug-in Fortran implementation of the neural networks for use in climate models.

Plain Language Summary

Heat can make its way into or out of sea ice via unpredictable air movements, known as turbulence, near the sea surface. In order to predict how quickly Arctic sea ice will melt in the future, we need to know how much heat the turbulence can transport in different weather conditions. Traditionally, turbulence calculations have been performed using sophisticated mathematical formulae from physics. In this study, we test an alternative method for predicting turbulent heat exchange: a computer algorithm known as an artificial neural network. By showing turbulence data, measured in the Arctic during previous scientific expeditions, to the network, it can be “trained” to make predictions in a process known as machine learning. We compare turbulence measurements, taken above sea ice in the recent MOSAiC expedition, with predictions from trained neural networks. We find that the neural networks are better than the traditional physics at predicting what the scientists at MOSAiC observed. The trained neural networks have been made publicly available so that they can be used by scientists for predicting climate change.
1 Introduction

The polar regions, in particular the Arctic, are on the front line of the climate crisis. In recent decades, the rate of surface warming in the Arctic has been two to four times higher than the global mean (Rantanen et al., 2022), a phenomenon known as Arctic amplification (e.g., Serreze & Francis, 2006; Serreze & Barry, 2011). Alongside rising temperatures have occurred losses of around 50 percent in both thickness and extent of Arctic sea ice at the end of summer since satellite records began (Gascard et al., 2019). The rate of Arctic sea ice loss in the coming decades remains highly uncertain (Bonan, Lehner, & Holland, 2021; Bonan, Schneider, et al., 2021), however the consequences are expected to be severe: for local ecosystems (Kovacs et al., 2011; Post et al., 2013; Tynan, 2015); for indigenous peoples (Meier et al., 2014); and, potentially, for lower-latitude climate (Cohen et al., 2014; Jung et al., 2015; Cohen et al., 2020; Liu et al., 2022). Heat exchanges between sea ice and the atmosphere are a key driver of the Arctic amplification (e.g., Serreze et al., 2009; Lesins et al., 2012; Previdi et al., 2021) and determine the sea ice melting rate (e.g., Rothrock et al., 1999; Screen & Simmonds, 2010).

Turbulent exchanges of heat and momentum in the planetary boundary layer are not directly simulated in climate models, but are instead represented through parametrizations, typically bulk formulae based on the Monin-Obukhov Similarity Theory (MOST, Monin & Obukhov, 1954; Garratt, 1994). Such parametrizations are semi-empirical: although the MOST provides dimensionless relationships, their final forms cannot be determined without recourse to observational data (e.g., calibration of roughness models and stability functions). The polar boundary layer is influenced by the presence of sea ice and is characterized by high stability and often intermittent turbulence (e.g., Andreas, 1998). Polar-specific stability functions have been proposed (Grachev et al., 2007), as well as formulations of surface roughness (e.g., Andreas, 1987; Andreas, Persson, et al., 2010; Andreas, 2011). More recently, parametrizations have been developed that account for form drag arising from alternating sea ice floes and leads (e.g., Lüpkes et al., 2012; Lüpkes & Gryani, 2015; Elvidge et al., 2016). Use of polar-specific turbulence parametrizations has been found to reduce biases in atmospheric models (Renfrew et al., 2019; Elvidge et al., 2023), however adoption of these advanced parametrizations in climate models has until recently been limited. The historic scarcity of observations in the Arctic likely goes some way to explaining modelers’ caution, yet there are also longstanding unresolved prob-
lems with modeling even homogeneous stable boundary layers (e.g., the GABLS experiments, Cuxart et al., 2006; Svensson et al., 2011; Bosveld et al., 2014).

Outside the polar regions, where observations have historically been more readily available, machine learning (ML) has emerged in recent years as an alternative strategy for parametrizing boundary-layer processes (Pal & Sharma, 2021). The basic idea of the ML or data-driven approach is that, given sufficient observational data, statistical algorithms can be used to directly infer empirical relationships between quantities of interest, such as surface turbulent fluxes, and mean meteorological variables such as temperature, humidity, etc. Recent studies have found that ML parametrizations, based on artificial neural networks (ANNs), can predict surface turbulent fluxes measured at meteorological towers in extra-polar regions with greater accuracy than bulk algorithms based on the MOST (Leuflen & Schädler, 2019; McCandless et al., 2022; Wulfmeyer et al., 2022). These findings were extended to the Arctic by Cummins et al. (2023), who showed that, even with the relatively small volume of data collected in previous Arctic campaigns, it is nevertheless possible to train ANNs that can outperform a polar-specific bulk algorithm.

The present study is motivated by the recent publication of surface turbulent flux observations collected at the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC, Shupe et al., 2022). The MOSAiC dataset has greatly increased the total number of flux observations collected over the Arctic sea ice, and thus presents a significant opportunity for validation and calibration of polar turbulence parametrizations. In this paper, we employ the MOSAiC data first to validate the performance of the ANNs of Cummins et al. (2023), using a MOST-based bulk algorithm as a baseline. We then incorporate the MOSAiC data into the ANN training set to generate an improved set of flux parametrizations for use in polar conditions (see Code Availability Statement). The remainder of this paper is organized as follows. Section 2 briefly recaps the datasets used in Cummins et al. (2023) and introduces the new MOSAiC data. Section 3 describes the ML and bulk algorithm flux parametrizations used in this study and the statistical methods used to evaluate their performance. Section 4 presents the results. Conclusions and recommendations for climate modelers are given in Section 5.
2 Data

2.1 Pre-MOSAiC observational campaigns

Cummins et al. (2023) trained and validated ANN models using surface turbulent flux measurements from four observational campaigns conducted over Arctic sea ice: Surface Heat Budget of the Arctic Ocean (SHEBA, Andreas et al., 1999; Persson et al., 2002; Uttal et al., 2002); Aerosol-Cloud Coupling and Climate Interactions in the Arctic (ACCACIA, Elvidge et al., 2016); Arctic Cloud in Summer Experiment (ACSE, Sotiropoulou et al., 2016; Prytherch et al., 2017); and Arctic Ocean 2016 (AO16, Tjernström & Jakobsson, 2021; Srivastava et al., 2022). These datasets sample a range of seasons and meteorological conditions in the Arctic. The sea ice varies in concentration (between zero and one), as well as in its morphology. For example, the ice surrounding the year-long SHEBA camp was compact and snow-covered in winter (Andreas, Persson, et al., 2010), but littered with deep melt ponds and leads in summer (Andreas, Horst, et al., 2010). It should be noted that Cummins et al. (2023) omitted from the training set observations in ACCACIA that were collected at heights > 30 m above the surface. Surface turbulent fluxes in climate models are typically calculated much closer to the surface (e.g., CNRM-CM6-1, Voldoire et al., 2019). Satellite estimates of sea ice concentration were obtained from the National Snow and Ice Data Center (NSIDC, Meier et al., 2021). See the Data Availability Statement for information about how to access these datasets.

2.2 MOSAiC

For the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) expedition, the icebreaker RV Polarstern was frozen into the Arctic sea ice and drifted with it for most of a year between Oct 2019 and Oct 2020. The original ice floe, on which the MOSAiC camp was established in Oct 2019, exited into the North Atlantic in late July 2020. Polarstern then repositioned near the North Pole at a new ice floe for August and September 2020. Various scientific research sites were established on the ice surrounding the ship, in a fashion similar to SHEBA although on a larger scale. As part of MOSAiC, extensive measurements were taken of the Arctic atmospheric system (Shupe et al., 2022). Surface turbulent fluxes of momentum, sensible heat and latent heat were computed at multiple locations using eddy-covariance techniques together with high-frequency (sampling rates of 10-20 Hz) observations from ultrasonic anemometers. Turbulence mea-
surements were made at a meteorological tower with sensors at 2, 6 and 10 m above the initial snow/ice surface. Data from all three tower levels were used in this study. Flux measurements were also taken at 3.8 m at the three Atmospheric Surface Flux Stations (ASFS), analogous to the Portable Automated Mesonet (PAM) stations in SHEBA. Note that, due to accumulation and ablation of snow, the actual measurement heights varied over time. All MOSAiC data used in this study were taken from publicly available repositories and were subject to Level3.4 quality control (see Data Availability Statement).

3 Methods

Surface turbulent fluxes are typically computed in climate models through bulk algorithms using wind, temperature and humidity at the single model level closest to the surface (e.g., in the SURFEX module in CNRM-CM6-1, Voldoire et al., 2019). The MOST, or a simplified version thereof, may then be used to extrapolate the vertical profiles of meteorological variables in the surface layer (Geleyn, 1988). Flux parametrizations in this study have been developed as plug-in replacements for bulk algorithms and therefore expect similar inputs. For the MOSAiC ASFS data, the wind and temperature/humidity measurements were made at different heights above the snow/ice surface (3.86 and 2.13/1.84 m respectively). While this doesn’t preclude a direct application of the bulk approach (since the MOST does not require measurements of those variables at the same height), it means that some pre-processing is required before the ANNs of Cummins et al. (2023) can be used. A decision was made not to attempt an interpolation of the wind speed, which was measured at a single height. Instead, the temperature/humidity measurements were linearly extrapolated to 3.86 m, using the observed gradient between the surface and the measurement height. Surface specific humidity was computed from temperature and pressure using the meteolib Python library (see Code Availability Statement). More sophisticated alternatives include a logarithmic extrapolation, or one based on the full MOST. However, our own numerical tests, conducted using equivalent measurements at 2 and 6 m on the meteorological tower, found the linear extrapolation to outperform the logarithmic in a root-mean-square error (RMSE) sense. Using the MOST approach would naturally introduce a bias in favour of that methodology. Different sensors were also mounted at slightly different heights around the nominal height of each tower level. Taking the heights of different sensors as the measurement height was found to have a small impact.
on the accuracy of flux predictions (±10% RMSE). In the final analysis, it was decided
to use the nominal heights of the tower levels, corrected for snow thickness, which is con-
sistent with how Cummins et al. (2023) treated data from the meteorological tower of
the SHEBA campaign.

Cummins et al. (2023) developed ML flux parametrizations based on single-layer,
feed-forward ANNs with four nodes in the hidden layer. For a high-level introduction
to statistical modeling with neural networks, see Hastie et al. (2009) or Kuhn and John-
son (2013). These models are general-purpose non-linear functions (Hornik et al., 1989),
permitting a high degree of variable interaction, and containing 37 tuneable parameters.
Each ANN takes seven mean meteorological variables as inputs: the measurement height
z; windspeed u(z); potential temperatures θ(z), θs of the air and at the surface respec-
tively; specific humidities q(z), qs; and the sea ice concentration Ci, determined over a
25×25 km² domain. The models were trained on the pre-MOSAiC data using the nnet
library for the statistical programming language R (Venables & Ripley, 2002; R Core Team,
2021). A weight decay of λ = 0.01 was used for regularization and the networks were
fitted in ensembles of 100 models to reduce variability due to random parameter initial-
ization (Ripley, 1996). The fitted ANNs output turbulent fluxes of momentum u₂*, sen-
sible heat u*θ*, and latent heat u*qs*. Predicted fluxes are returned in kinematic units,
i.e. in the same units as the measured eddy covariances, and hence are written here in
terms of the MOST scaling parameters u*, θ*, q*.

The polar-specific bulk algorithm, used in this study as a baseline against which
to compare the ANNs, is the same as that described in Cummins et al. (2023). Over open
water, the iterative COARE 3.0 algorithm is used (Fairall et al., 2003; Edson et al., 2013),
with stability functions from Grachev et al. (2000) in unstable conditions and from Beljaars
and Holtslag (1991) in stable conditions. The COARE 3.0 algorithm has been well tested
over the years and is currently in use in large-scale climate models, including CNRM-
CM6-1. Bulk transfer coefficients are initialized using a non-iterative estimate of the sta-
bility (Grachev & Fairall, 1997). Over sea ice, the stability function from Grachev et al.
(2007) is used in stable conditions, as well as the scalar roughness model of Andreas (1987)
and the aerodynamic roughness model of Andreas, Persson, et al. (2010). For partial sea
ice concentrations, we use the mosaic approach (e.g., Vihma, 1995), whereby we take a
weighted average of fluxes computed over open water and over sea ice, with the weight-
ing given by the sea ice concentration. The mosaic approach is currently used in CNRM-
An additional form drag contribution is included when computing the momentum flux, to account for the influence of intermittent sea ice coverage (Lügkes & Gryanik, 2015). Intermittent ice coverage is associated with vertical ice surfaces that tend to increase turbulence. This bulk algorithm is available for download as a Python library (see Code Availability Statement). Compared against estimates from unmodified COARE 3.0, momentum flux estimates from our bulk algorithm have lower RMSE at the MOSAiC sites (up to a 16% reduction). The polar-specific components have less impact on the heat fluxes: there is a 99% correlation between our heat fluxes and those from COARE 3.0. The results of our comparison with ML in Section 4 are robust to the use / non-use of polar-specific components in the bulk algorithm.

In Cummins et al. (2023), the ML and bulk algorithm flux parametrizations were tested using a campaign-wise cross-validation scheme. Each campaign (or measurement site in the case of SHEBA) was left out of the training set in turn and the trained models validated on that campaign. Flux predictions from the two methods, together with measured eddy covariances, were used to compute performance metrics, such as RMSE, mean absolute error (MAE) and Pearson correlation. Since the MOSAiC data were not involved in the calibration of either parametrization, they constitute an independent test set and are therefore ideal for model validation and comparison. Mean meteorological variables, measured at each of the MOSAiC sites, were supplied as input variables and predicted fluxes calculated. In addition to these truly out-of-sample predictions, further flux estimates were obtained from ANNs fitted to MOSAiC-augmented training sets: for each site in MOSAiC, an ANN model was fitted to a training set comprising the pre-MOSAiC data plus all MOSAiC data not observed at that site. Iterating over the MOSAiC sites then gives a complete set of out-of-sample predictions, which allows us to quantify any gains in predictive power obtained from the MOSAiC data.

4 Results

Performance metrics, computed for the bulk algorithm and ANN parametrizations at each of the MOSAiC sites, are given in Table 1. Figures 1-3 show two-dimensional histograms of predicted fluxes against measured eddy covariances at each site. Note that results at the meteorological tower do not differ qualitatively between tower levels in terms of patterns/biases, however there is a small dependence of predictive accuracy on measurement height. Specifically, both the bulk algorithm and neural network methods have
Table 1. Predictive performance of neural network (nnet/nnet+), and Monin-Obukhov (bulk) flux parametrizations at the MOSAiC sites in kinematic units. The nnet+ columns show results for ANNs trained using MOSAiC-augmented data. Boldface indicates a better score in one of root-mean-square error (RMSE), mean absolute error (MAE) or Pearson correlation. Using bootstrapping, all score differences were found to be statistically significant at the five-percent level (Davison & Hinkley, 1997). Note that direct measurements of $u_\star q_\star$ are not available at ASFS40.

<table>
<thead>
<tr>
<th>site</th>
<th>n</th>
<th>RMSE</th>
<th>MAE</th>
<th>corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>bulk</td>
<td>nnet</td>
<td>nnet+</td>
</tr>
<tr>
<td>$u_\star^2$</td>
<td>ASFS30</td>
<td>22161</td>
<td>0.039</td>
<td>0.035</td>
</tr>
<tr>
<td>ASFS40</td>
<td>18498</td>
<td>0.035</td>
<td>0.034</td>
<td>0.029</td>
</tr>
<tr>
<td>ASFS50</td>
<td>18871</td>
<td>0.031</td>
<td>0.027</td>
<td>0.025</td>
</tr>
<tr>
<td>met. tower</td>
<td>66777</td>
<td>0.049</td>
<td>0.047</td>
<td>0.044</td>
</tr>
<tr>
<td>$u_\star \theta_\star$</td>
<td>ASFS30</td>
<td>22161</td>
<td>0.0091</td>
<td>0.0057</td>
</tr>
<tr>
<td>ASFS40</td>
<td>18498</td>
<td>0.0113</td>
<td>0.0071</td>
<td>0.0058</td>
</tr>
<tr>
<td>ASFS50</td>
<td>18871</td>
<td>0.0056</td>
<td>0.0045</td>
<td>0.0049</td>
</tr>
<tr>
<td>met. tower</td>
<td>66777</td>
<td>0.0073</td>
<td>0.0066</td>
<td>0.0062</td>
</tr>
<tr>
<td>$u_\star q_\star$</td>
<td>ASFS30</td>
<td>2692</td>
<td>3.20E-06</td>
<td>8.43E-07</td>
</tr>
<tr>
<td>ASFS50</td>
<td>3436</td>
<td>9.88E-07</td>
<td>8.93E-07</td>
<td>8.50E-07</td>
</tr>
<tr>
<td>met. tower</td>
<td>33185</td>
<td>5.30E-07</td>
<td>5.94E-07</td>
<td>5.87E-07</td>
</tr>
</tbody>
</table>
slightly lower RMSE (∼10%) when applied at 10 m compared with 2 m. This is as ex-
pected: the 10 m gradients of the meteorological variables are larger than the correspond-
ing 2 m gradients, so if the measurement errors at the different levels are similar in mag-
nitude then the 10 m gradients should have lower relative error. Any inaccuracies in the
estimated measurement heights should also be proportionally smaller at 10 m. Overall,
the results are encouraging, with the ANN parametrizations consistently delivering per-
formance improvements over the bulk algorithm.

Both methods produce similar estimates of the momentum flux $u_2^*$ and the resid-
ual plots in Figure 1 share common features, such as a conservative bias (systematic un-
derprediction of larger fluxes). However, the ANNs achieve a lower RMSE at all the MO-
SAiC sites: a result which is robust under bootstrap resampling (Davison & Hinkley, 1997).
The tendency of the ANNs to underpredict the magnitude of large and extreme fluxes
was noted by Cummins et al. (2023) and is a known property of the models. In short,
the ANNs have an inbuilt reluctance to extrapolate when faced with a combination of
inputs not seen in training. That the bulk algorithm also underpredicts $u_2^*$ is unexpected
and warrants investigation (see final paragraph of this section). Augmenting the ANN
training set with data from MOSAiC further reduces the RMSE of the ANNs at all sites,
as well as producing a visible attenuation of the conservative bias for larger fluxes. This
result indicates that meteorological conditions conducive to large $u_2^*$ were consistent across
the MOSAiC measurement sites.

The ANN parametrization strongly outperforms the bulk algorithm as an estima-
tor of the sensible heat flux $u_*\theta_*$, with RMSE 10-40 percent lower across the sites. It can
be seen from Figure 2 that the improvements over the bulk are particularly apparent at
the ASFS30 and ASFS40 stations. As was the case for $u_2^*$, the prediction errors of the
two $u_*\theta_*$ parametrizations share common features, including some clearly non-random
deviations from the line $y = x$. In particular, there is a long tail in the panels for the
met. tower in Figure 2. The tail is comprised of large negative fluxes whose magnitude
was underestimated by both the bulk and ANN parametrizations. The underestimation
occurred because, despite the large fluxes, the average temperature gradient over these
10-minute sampling periods was small. This highlights a limitation of the time-averaging
approach to measuring fluxes and indicates that, potentially, new physics may be required
to explain this phenomenon. It is possible that these large values in the direct measure-
ments are the result of non-stationarity, i.e., that not all of the flux is really turbulence.
Figure 1. Predicted momentum fluxes $u^2$ at the MOSAiC sites in kinematic units plotted against observed eddy covariances. To the left are estimates obtained from a polar-specific bulk algorithm based on the Monin-Obukhov Similarity Theory; in the centre, estimates from the neural networks of Cummins et al. (2023); to the right, estimates from neural networks trained using MOSAiC-augmented data. The diagonal line $y = x$ would represent a perfect fit.
Figure 2. Predicted sensible heat fluxes $u_\star, \theta_\star$, at the MOSAiC sites in kinematic units plotted against observed eddy covariances. See Figure 1 caption for details.
At the tower, there is also evidence of an inverse phenomenon, where large gradients are observed but the corresponding fluxes are small. Including MOSAiC data from other sites in the ANN training set produces clear improvements at three of the four sites, with several systematic features in the residuals disappearing. The predictions at the ASFS50 site, however, became worse. Prediction errors at ASFS50 are the lowest for $u^*_{\text{th}}$ and $u^*_{\text{v}}$, so it doesn’t necessarily follow that the site is at fault. It is entirely possible that there are locally varying factors, not included in the set of input variables, which affect the fluxes (see final paragraph of this section). Identifying such variables has the potential to deliver further performance gains.

Latent heat flux $u^*_q$ is by far the most difficult of the three fluxes to predict: $u^*_q$ was generally small in magnitude at the MOSAiC measurement sites, suggesting a low signal-to-noise ratio. The ANNs are also disadvantaged here by a small training set from previous campaigns, comprised mainly of very small fluxes (Cummins et al., 2023). Results for $u^*_q$ are therefore unsurprising: to the extent that the measured fluxes are small in magnitude, the ANNs perform well. For larger fluxes, the ANNs exhibit a strong conservative bias. Conversely, the bulk algorithm tends to overpredict the magnitude of $u^*_q$. It is because of these contrasting biases that the bulk achieves a higher correlation at the ASFS30 and ASFS50 stations, while at the same time the ANNs give RMSE reductions at those sites of about 70 and 10 percent respectively. At the MOSAiC tower, the bulk algorithm performs better, achieving a 10-percent lower RMSE. The ANNs trained on the MOSAiC-augmented data perform better at the ASFS50 site and the tower, but slightly worse at ASFS30. From Figure 3 it can be seen that the underestimation bias, while still present, is improved by training with MOSAiC data.

It should be noted that the biases, visible in the MOSAiC-augmented ANN predictions in Figures 1-3, should be further reduced by the next step, which is to incorporate all the MOSAiC data in the ANN training set. As more observations become available, we would expect the as-yet-unsampled regions of the input space to diminish, along with the associated biases. That is not to say that all biases can be resolved through more training data. As mentioned above, omission of important predictor variables has the potential to induce biases that will persist regardless of the volume of training data. For example, the upwards and downwards radiation terms are known to contribute significant explanatory power (e.g., Wulfmeyer et al., 2022). These radiation terms are nevertheless unsuitable for use as parametrization inputs, because the radiative fluxes in GCMs...
Figure 3. Predicted latent heat fluxes $u_* q_*$ at the MOSAiC sites in kinematic units plotted against observed eddy covariances. See Figure 1 caption for details.
are themselves the output of complex parametrizations with their own errors and uncertainties. The surface characteristics may also be an important missing variable. In MOSAiC, the winter sea ice may have been generally aerodynamically rougher than that seen in the SHEBA campaign. This could potentially explain the underestimation of $u^*_3$ by both the ANN and bulk algorithm parametrizations (see Figure 1).

5 Conclusions

Accurate representation in climate models of turbulent heat exchanges between the surface and atmosphere in polar regions is essential for constraining predictions of future climate change, locally and potentially globally. Surface turbulent fluxes in the polar boundary layer are currently parametrized using the traditional MOST, although alternative ML parametrizations based on ANNs have recently been proposed (Cummins et al., 2023). The wealth of new flux observations collected in the Arctic during the MOSAiC campaign has provided an excellent opportunity to validate and calibrate these alternative parametrizations.

In this study, the MOSAiC data have been used to validate ANN parametrizations of momentum, sensible heat and latent heat fluxes, that were originally trained on data from previous Arctic campaigns. The ANNs have been found to generalize well to the new data, particularly for momentum and sensible heat, yielding substantial reductions in error metrics such as RMSE when compared against a polar-specific bulk algorithm based on the MOST. Although the ANNs performed well at predicting small latent heat fluxes, limitations of the training data resulted in systematic underprediction of larger fluxes.

The ANN parametrizations, developed in Cummins et al. (2023) and validated in this study, have been recalibrated on an augmented training dataset that incorporates the observations from MOSAiC. These updated parametrizations have been implemented as a Fortran subroutine, suitable for deployment in climate models as a plug-in replacement for bulk algorithms (see Code Availability Statement). An important next step will be to perform sensitivity studies with these new parametrizations in a climate model. In this way, the implications for the polar atmosphere and melting of Arctic sea ice can be assessed.
Code Availability Statement

Tools to reproduce the results presented in this study are publicly available at the following repositories:

- [https://doi.org/10.5281/zenodo.8207293](https://doi.org/10.5281/zenodo.8207293) This Python library provides functions to compute transfer coefficients and related variables (zeta, stability functions, aerodynamic and scalar roughness etc.), as well as to apply the bulk algorithm parametrizations used in this study.

- [https://doi.org/10.5281/zenodo.8207302](https://doi.org/10.5281/zenodo.8207302) This Python library provides functions to estimate meteorological parameters (humidity, latent heat as a function of temperature etc.).

- [https://doi.org/10.5281/zenodo.8207288](https://doi.org/10.5281/zenodo.8207288) This Fortran subroutine implements the neural network flux parametrizations developed in this study. The networks have been trained on all available datasets, including MOSAiC.

Data Availability Statement

MOSAiC campaign sites

The MOSAiC data used in this study are available from the National Science Foundation Arctic Data Center: met. tower ([https://doi.org/10.18739/A2PV6B83F](https://doi.org/10.18739/A2PV6B83F), Cox, Gallagher, Shupe, Persson, Blomquist, et al., 2023); ASFS30 ([https://doi.org/10.18739/A2FF3M18K](https://doi.org/10.18739/A2FF3M18K), Cox, Gallagher, Shupe, Persson, Grachev, et al., 2023a); ASFS40 ([https://doi.org/10.18739/A25X25F0P](https://doi.org/10.18739/A25X25F0P), Cox, Gallagher, Shupe, Persson, Grachev, et al., 2023b), ASFS50 ([https://doi.org/10.18739/A2XD0R00S](https://doi.org/10.18739/A2XD0R00S), Cox, Gallagher, Shupe, Persson, Grachev, et al., 2023c).

Pre-MOSAiC observational campaigns

The SHEBA data are available from the UCAR Earth Observing Laboratory: met. tower ([https://doi.org/10.5065/D65H7DNS](https://doi.org/10.5065/D65H7DNS), Andreas et al., 2007); PAM stations ([https://doi.org/10.5065/D6ZC8170](https://doi.org/10.5065/D6ZC8170), Andreas et al., 2012). The ACCACIA flight data are available from the CEDA archive: [https://doi.org/10.5285/0844186db1ba9e20319a2560f8d61651](https://doi.org/10.5285/0844186db1ba9e20319a2560f8d61651) (MASIN); [https://catalogue.ceda.ac.uk/uuid/c064b0c150274a1c63573f392e](https://catalogue.ceda.ac.uk/uuid/c064b0c150274a1c63573f392e) (FAAM). The ACSE cruise data are available from the CEDA archive ([https://doi.org/](https://doi.org/)}
Acknowledgments

This work was supported by a national funding by the Agence Nationale de la Recherche within the framework of the Investissement d’Avenir program under the ANR-17-MPGA-0003 reference. This article has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101003826 via project CRIceS (Climate Relevant interactions and feedbacks: the key role of sea ice and Snow in the polar and global climate system).

Data used in this manuscript were produced as part of the international Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) with the tag MOSAiC20192020. We thank all persons involved in the expedition of the Research Vessel Polarstern during MOSAiC in 2019-2020 (AWI_PS122_00) as listed in Nixdorf et al. (2021). The MOSAiC observations received support from the US National Science Foundation Office of Polar Programs (OPP-1724551); the NOAA Physical Sciences Laboratory; and NOAA’s Global Ocean Monitoring and Observing Program (FundRef https://doi.org/10.13039/100018302).

This is a contribution to the Year of Polar Prediction (YOPP), a flagship activity of the Polar Prediction Project (PPP), initiated by the World Weather Research Programme (WWRP) of the World Meteorological Organisation (WMO). We acknowledge the WMO WWRP for its role in coordinating this international research activity.

Neural network ensembles were fitted using the R package caret (Kuhn & Johnson, 2013), which itself depends on the R package nnet (Venables & Ripley, 2002) to fit the underlying models. Bootstrapping of model performance metrics was performed using the R package boot (Canty & Ripley, 2022).

References


Brooks, I. M., Prytherch, J., & Srivastava, P. (2022a). *CANDIFLOS: Surface fluxes from ACSE measurement campaign on icebreaker Oden, 2014* [Dataset]. NERC EDS Centre for Environmental Data Analysis. doi: 10.5285/C6F1B1FF16F8407386E2D643BC5B916A


Canty, A., & Ripley, B. D. (2022, November). *Boot: Bootstrap Functions (Originally by Angelo Canty for S).*


Cox, C., Gallagher, M., Shupe, M., Persson, O., Grachev, A., Solomon, A., . . . Ut- tal, T. (2023a). *Atmospheric Surface Flux Station #30 measurements (Level 3*
Manuscript submitted to Geophysical Research Letters

Final), Multidisciplinary Drifting Observatory for the Study of Arctic Climate (MOSAiC), central Arctic, October 2019 - September 2020. [Dataset]. NSF Arctic Data Center. doi: 10.18739/A2FF3M18K


change over the marginal ice zone and recommendations for its parametri-

Atmospheric Chemistry and Physics, 16(3), 1545–1563. doi:
10.5194/acp-16-1545-2016

Fairall, C. W., Bradley, E. F., Hare, J. E., Grachev, A. A., & Edson, J. B. (2003, February). Bulk Parameterization of Air–Sea Fluxes: Updates and Verifica-
tion for the COARE Algorithm. Journal of Climate, 16(4), 571–591. doi:

Garratt, J. R. (1994). The Atmospheric Boundary Layer. Cambridge, UK: Cam-
bridge University Press.

Gascard, J.-C., Zhang, J., & Rafizadeh, M. (2019, January). Rapid decline of Ar-
tic sea ice volume: Causes and consequences. The Cryosphere Discussions, 1–
29. doi: 10.5194/tc-2019-2

model levels to the height of measurement. Tellus A, 40A(4), 347–351. doi: 10
.1111/j.1600-0870.1988.tb00352.x

(2007, September). SHEBA flux–profile relationships in the stable atmo-
spheric boundary layer. Boundary-Layer Meteorology, 124(3), 315–333. doi:
10.1007/s10546-007-9177-6

Stability Parameter on the Bulk Richardson Number over the Ocean.
Journal of Applied Meteorology and Climatology, 36(4), 406–414. doi:
10.1175/1520-0450(1997)036⟨⟨0406:DOTMOS⟩⟩2.0.CO;2

Constants Revisited. Boundary-Layer Meteorology, 94(3), 495–515. doi: 10
.1023/A:1002452529672

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Neural Networks. In T. Hastie,
R. Tibshirani, & J. Friedman (Eds.), The Elements of Statistical Learning:
Data Mining, Inference, and Prediction (pp. 389–416). New York, NY:
Springer. doi: 10.1007/978-0-387-84858-7\_11

networks are universal approximators. Neural Networks, 2(5), 359–366. doi: 10
.1016/0893-6080(89)90020-8


Surface turbulent fluxes from the MOSAiC campaign
predicted by machine learning

Donald P. Cummins¹, Virginie Guemas¹, Christopher J. Cox², Michael R.
Gallagher²,³, Matthew D. Shupe²,³

¹CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France
²National Oceanic and Atmospheric Administration, Physical Sciences Laboratory, Boulder, CO, USA
³Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA

Key Points:

• Neural networks trained on previous Arctic campaigns predict surface turbulent fluxes from MOSAiC more accurately than bulk methods.
• Updated parametrizations using the MOSAiC data have been developed and implemented in Fortran for deployment in climate models.

Corresponding author: Donald P. Cummins, donald.cummins@meteo.fr
Abstract

Reliable boundary-layer turbulence parametrizations for polar conditions are needed to reduce uncertainty in projections of Arctic sea ice melting rate and its potential global repercussions. Surface turbulent fluxes of sensible and latent heat are typically represented in climate models using bulk formulae based on the Monin-Obukhov Similarity Theory (MOST), sometimes finely tuned to high stability conditions and the potential presence of sea ice. In this study, we test the performance of new, machine-learning (ML) flux parametrizations, using an advanced polar-specific bulk algorithm as a baseline. Neural networks, trained on observations from previous Arctic campaigns, are used to predict surface turbulent fluxes measured over sea ice as part of the recent MOSAiC expedition. The ML parametrizations outperform the bulk at the MOSAiC sites, with RMSE reductions of up to 70 percent. We provide a plug-in Fortran implementation of the neural networks for use in climate models.

Plain Language Summary

Heat can make its way into or out of sea ice via unpredictable air movements, known as turbulence, near the sea surface. In order to predict how quickly Arctic sea ice will melt in the future, we need to know how much heat the turbulence can transport in different weather conditions. Traditionally, turbulence calculations have been performed using sophisticated mathematical formulae from physics. In this study, we test an alternative method for predicting turbulent heat exchange: a computer algorithm known as an artificial neural network. By showing turbulence data, measured in the Arctic during previous scientific expeditions, to the network, it can be “trained” to make predictions in a process known as machine learning. We compare turbulence measurements, taken above sea ice in the recent MOSAiC expedition, with predictions from trained neural networks. We find that the neural networks are better than the traditional physics at predicting what the scientists at MOSAiC observed. The trained neural networks have been made publicly available so that they can be used by scientists for predicting climate change.
1 Introduction

The polar regions, in particular the Arctic, are on the front line of the climate crisis. In recent decades, the rate of surface warming in the Arctic has been two to four times higher than the global mean (Rantanen et al., 2022), a phenomenon known as Arctic amplification (e.g., Serreze & Francis, 2006; Serreze & Barry, 2011). Alongside rising temperatures have occurred losses of around 50 percent in both thickness and extent of Arctic sea ice at the end of summer since satellite records began (Gascard et al., 2019). The rate of Arctic sea ice loss in the coming decades remains highly uncertain (Bonan, Lehner, & Holland, 2021; Bonan, Schneider, et al., 2021), however the consequences are expected to be severe: for local ecosystems (Kovacs et al., 2011; Post et al., 2013; Tynan, 2015); for indigenous peoples (Meier et al., 2014); and, potentially, for lower-latitude climate (Cohen et al., 2014; Jung et al., 2015; Cohen et al., 2020; Liu et al., 2022). Heat exchanges between sea ice and the atmosphere are a key driver of the Arctic amplification (e.g., Serreze et al., 2009; Lesins et al., 2012; Previdi et al., 2021) and determine the sea ice melting rate (e.g., Rothrock et al., 1999; Screen & Simmonds, 2010).

Turbulent exchanges of heat and momentum in the planetary boundary layer are not directly simulated in climate models, but are instead represented through parametrizations, typically bulk formulae based on the Monin-Obukhov Similarity Theory (MOST, Monin & Obukhov, 1954; Garratt, 1994). Such parametrizations are semi-empirical: although the MOST provides dimensionless relationships, their final forms cannot be determined without recourse to observational data (e.g., calibration of roughness models and stability functions). The polar boundary layer is influenced by the presence of sea ice and is characterized by high stability and often intermittent turbulence (e.g., Andreas, 1998). Polar-specific stability functions have been proposed (Grachev et al., 2007), as well as formulations of surface roughness (e.g., Andreas, 1987; Andreas, Persson, et al., 2010; Andreas, 2011). More recently, parametrizations have been developed that account for form drag arising from alternating sea ice floes and leads (e.g., Lüpkes et al., 2012; Lüpkes & Gryanik, 2015; Elvidge et al., 2016). Use of polar-specific turbulence parametrizations has been found to reduce biases in atmospheric models (Renfrew et al., 2019; Elvidge et al., 2023), however adoption of these advanced parametrizations in climate models has until recently been limited. The historic scarcity of observations in the Arctic likely goes some way to explaining modelers’ caution, yet there are also longstanding unresolved prob-
lems with modeling even homogeneous stable boundary layers (e.g., the GABLS experiments, Cuxart et al., 2006; Svensson et al., 2011; Bosveld et al., 2014).

Outside the polar regions, where observations have historically been more readily available, machine learning (ML) has emerged in recent years as an alternative strategy for parametrizing boundary-layer processes (Pal & Sharma, 2021). The basic idea of the ML or data-driven approach is that, given sufficient observational data, statistical algorithms can be used to directly infer empirical relationships between quantities of interest, such as surface turbulent fluxes, and mean meteorological variables such as temperature, humidity, etc. Recent studies have found that ML parametrizations, based on artificial neural networks (ANNs), can predict surface turbulent fluxes measured at meteorological towers in extra-polar regions with greater accuracy than bulk algorithms based on the MOST (Leufen & Schälder, 2019; McCandless et al., 2022; Wulfmeyer et al., 2022). These findings were extended to the Arctic by Cummins et al. (2023), who showed that, even with the relatively small volume of data collected in previous Arctic campaigns, it is nevertheless possible to train ANNs that can outperform a polar-specific bulk algorithm.

The present study is motivated by the recent publication of surface turbulent flux observations collected at the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC, Shupe et al., 2022). The MOSAiC dataset has greatly increased the total number of flux observations collected over the Arctic sea ice, and thus presents a significant opportunity for validation and calibration of polar turbulence parametrizations. In this paper, we employ the MOSAiC data first to validate the performance of the ANNs of Cummins et al. (2023), using a MOST-based bulk algorithm as a baseline. We then incorporate the MOSAiC data into the ANN training set to generate an improved set of flux parametrizations for use in polar conditions (see Code Availability Statement). The remainder of this paper is organized as follows. Section 2 briefly recaps the datasets used in Cummins et al. (2023) and introduces the new MOSAiC data. Section 3 describes the ML and bulk algorithm flux parametrizations used in this study and the statistical methods used to evaluate their performance. Section 4 presents the results. Conclusions and recommendations for climate modelers are given in Section 5.
2 Data

2.1 Pre-MOSAiC observational campaigns

Cummins et al. (2023) trained and validated ANN models using surface turbulent flux measurements from four observational campaigns conducted over Arctic sea ice: Surface Heat Budget of the Arctic Ocean (SHEBA, Andreas et al., 1999; Persson et al., 2002; Uttal et al., 2002); Aerosol-Cloud Coupling and Climate Interactions in the Arctic (ACCACIA, Elvidge et al., 2016); Arctic Cloud in Summer Experiment (ACSE, Sotiropoulou et al., 2016; Prytherch et al., 2017); and Arctic Ocean 2016 (AO16, Tjernström & Jakobsson, 2021; Srivastava et al., 2022). These datasets sample a range of seasons and meteorological conditions in the Arctic. The sea ice varies in concentration (between zero and one), as well as in its morphology. For example, the ice surrounding the year-long SHEBA camp was compact and snow-covered in winter (Andreas, Persson, et al., 2010), but littered with deep melt ponds and leads in summer (Andreas, Horst, et al., 2010). It should be noted that Cummins et al. (2023) omitted from the training set observations in ACCACIA that were collected at heights > 30 m above the surface. Surface turbulent fluxes in climate models are typically calculated much closer to the surface (e.g., CNRM-CM6-1, Voldoire et al., 2019). Satellite estimates of sea ice concentration were obtained from the National Snow and Ice Data Center (NSIDC, Meier et al., 2021). See the Data Availability Statement for information about how to access these datasets.

2.2 MOSAiC

For the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) expedition, the icebreaker RV Polarstern was frozen into the Arctic sea ice and drifted with it for most of a year between Oct 2019 and Oct 2020. The original ice floe, on which the MOSAiC camp was established in Oct 2019, exited into the North Atlantic in late July 2020. Polarstern then repositioned near the North Pole at a new ice floe for August and September 2020. Various scientific research sites were established on the ice surrounding the ship, in a fashion similar to SHEBA although on a larger scale. As part of MOSAiC, extensive measurements were taken of the Arctic atmospheric system (Shupe et al., 2022). Surface turbulent fluxes of momentum, sensible heat and latent heat were computed at multiple locations using eddy-covariance techniques together with high-frequency (sampling rates of 10-20 Hz) observations from ultrasonic anemometers. Turbulence mea-
measurements were made at a meteorological tower with sensors at 2, 6 and 10 m above the initial snow/ice surface. Data from all three tower levels were used in this study. Flux measurements were also taken at 3.8 m at the three Atmospheric Surface Flux Stations (ASFS), analogous to the Portable Automated Mesonet (PAM) stations in SHEBA. Note that, due to accumulation and ablation of snow, the actual measurement heights varied over time. All MOSAiC data used in this study were taken from publicly available repositories and were subject to Level3.4 quality control (see Data Availability Statement).

3 Methods

Surface turbulent fluxes are typically computed in climate models through bulk algorithms using wind, temperature and humidity at the single model level closest to the surface (e.g., in the SURFEX module in CNRM-CM6-1, Voldoire et al., 2019). The MOST, or a simplified version thereof, may then be used to extrapolate the vertical profiles of meteorological variables in the surface layer (Geleyn, 1988). Flux parametrizations in this study have been developed as plug-in replacements for bulk algorithms and therefore expect similar inputs. For the MOSAiC ASFS data, the wind and temperature/humidity measurements were made at different heights above the snow/ice surface (3.86 and 2.13/1.84 m respectively). While this doesn’t preclude a direct application of the bulk approach (since the MOST does not require measurements of those variables at the same height), it means that some pre-processing is required before the ANNs of Cummins et al. (2023) can be used. A decision was made not to attempt an interpolation of the wind speed, which was measured at a single height. Instead, the temperature/humidity measurements were linearly extrapolated to 3.86 m, using the observed gradient between the surface and the measurement height. Surface specific humidity was computed from temperature and pressure using the meteolib Python library (see Code Availability Statement). More sophisticated alternatives include a logarithmic extrapolation, or one based on the full MOST. However, our own numerical tests, conducted using equivalent measurements at 2 and 6 m on the meteorological tower, found the linear extrapolation to outperform the logarithmic in a root-mean-square error (RMSE) sense. Using the MOST approach would naturally introduce a bias in favour of that methodology. Different sensors were also mounted at slightly different heights around the nominal height of each tower level. Taking the heights of different sensors as the measurement height was found to have a small impact.
on the accuracy of flux predictions (±10% RMSE). In the final analysis, it was decided
to use the nominal heights of the tower levels, corrected for snow thickness, which is con-
sistent with how Cummins et al. (2023) treated data from the meteorological tower of
the SHEBA campaign.

Cummins et al. (2023) developed ML flux parametrizations based on single-layer,
feed-forward ANNs with four nodes in the hidden layer. For a high-level introduction
to statistical modeling with neural networks, see Hastie et al. (2009) or Kuhn and John-
son (2013). These models are general-purpose non-linear functions (Hornik et al., 1989),
permitting a high degree of variable interaction, and containing 37 tuneable parameters.
Each ANN takes seven mean meteorological variables as inputs: the measurement height
$z$; windspeed $u(z)$; potential temperatures $\theta(z)$, $\theta_s$ of the air and at the surface respec-
tively; specific humidities $q(z)$, $q_s$; and the sea ice concentration $C_i$, determined over a
25×25 km$^2$ domain. The models were trained on the pre-MOSAiC data using the \textit{nnet}
library for the statistical programming language R (Venables & Ripley, 2002; R Core Team,
2021). A weight decay of $\lambda = 0.01$ was used for regularization and the networks were
fitted in ensembles of 100 models to reduce variability due to random parameter initial-
ization (Ripley, 1996). The fitted ANNs output turbulent fluxes of momentum $u^*$, sen-
sible heat $u^*\theta_s$ and latent heat $u^*q_s$. Predicted fluxes are returned in kinematic units,
i.e. in the same units as the measured eddy covariances, and hence are written here in
terms of the MOST scaling parameters $u^*$, $\theta^*$, $q^*$.

The polar-specific bulk algorithm, used in this study as a baseline against which
to compare the ANNs, is the same as that described in Cummins et al. (2023). Over open
water, the iterative COARE 3.0 algorithm is used (Fairall et al., 2003; Edson et al., 2013),
with stability functions from Grachev et al. (2000) in unstable conditions and from Beljaars
and Holtslag (1991) in stable conditions. The COARE 3.0 algorithm has been well tested
over the years and is currently in use in large-scale climate models, including CNRM-
CM6-1. Bulk transfer coefficients are initialized using a non-iterative estimate of the sta-
bility (Grachev & Fairall, 1997). Over sea ice, the stability function from Grachev et al.
(2007) is used in stable conditions, as well as the scalar roughness model of Andreas (1987)
and the aerodynamic roughness model of Andreas, Persson, et al. (2010). For partial sea
ice concentrations, we use the \textit{mosaic} approach (e.g., Vihma, 1995), whereby we take a
weighted average of fluxes computed over open water and over sea ice, with the weight-
ing given by the sea ice concentration. The mosaic approach is currently used in CNRM-
CM6-1. An additional form drag contribution is included when computing the momentum flux, to account for the influence of intermittent sea ice coverage (Lüpkes & Gryanik, 2015). Intermittent ice coverage is associated with vertical ice surfaces that tend to increase turbulence. This bulk algorithm is available for download as a Python library (see Code Availability Statement). Compared against estimates from unmodified COARE 3.0, momentum flux estimates from our bulk algorithm have lower RMSE at the MOSAiC sites (up to a 16% reduction). The polar-specific components have less impact on the heat fluxes: there is a 99% correlation between our heat fluxes and those from COARE 3.0. The results of our comparison with ML in Section 4 are robust to the use / non-use of polar-specific components in the bulk algorithm.

In Cummins et al. (2023), the ML and bulk algorithm flux parametrizations were tested using a campaign-wise cross-validation scheme. Each campaign (or measurement site in the case of SHEBA) was left out of the training set in turn and the trained models validated on that campaign. Flux predictions from the two methods, together with measured eddy covariances, were used to compute performance metrics, such as RMSE, mean absolute error (MAE) and Pearson correlation. Since the MOSAiC data were not involved in the calibration of either parametrization, they constitute an independent test set and are therefore ideal for model validation and comparison. Mean meteorological variables, measured at each of the MOSAiC sites, were supplied as input variables and predicted fluxes calculated. In addition to these truly out-of-sample predictions, further flux estimates were obtained from ANNs fitted to MOSAiC-augmented training sets: for each site in MOSAiC, an ANN model was fitted to a training set comprising the pre-MOSAiC data plus all MOSAiC data not observed at that site. Iterating over the MOSAiC sites then gives a complete set of out-of-sample predictions, which allows us to quantify any gains in predictive power obtained from the MOSAiC data.

4 Results

Performance metrics, computed for the bulk algorithm and ANN parametrizations at each of the MOSAiC sites, are given in Table 1. Figures 1-3 show two-dimensional histograms of predicted fluxes against measured eddy covariances at each site. Note that results at the meteorological tower do not differ qualitatively between tower levels in terms of patterns/biases, however there is a small dependence of predictive accuracy on measurement height. Specifically, both the bulk algorithm and neural network methods have
Table 1. Predictive performance of neural network (nnet/nnet+), and Monin-Obukhov (bulk) flux parametrizations at the MOSAiC sites in kinematic units. The nnet+ columns show results for ANNs trained using MOSAiC-augmented data. Boldface indicates a better score in one of root-mean-square error (RMSE), mean absolute error (MAE) or Pearson correlation. Using bootstrapping, all score differences were found to be statistically significant at the five-percent level (Davison & Hinkley, 1997). Note that direct measurements of $u_\star q_\star$ are not available at ASFS40.

<table>
<thead>
<tr>
<th>site</th>
<th>n</th>
<th>RMSE</th>
<th>MAE</th>
<th>corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>bulk nnet nnet+</td>
<td>bulk nnet nnet+</td>
<td>bulk nnet nnet+</td>
</tr>
<tr>
<td>$u_\star^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASFS30</td>
<td>22161</td>
<td>0.039 0.035 0.033</td>
<td>0.021 0.018 0.019</td>
<td>0.92 0.92 0.92</td>
</tr>
<tr>
<td>ASFS40</td>
<td>18498</td>
<td>0.035 0.034 0.029</td>
<td>0.020 0.019 0.017</td>
<td>0.94 0.93 0.93</td>
</tr>
<tr>
<td>ASFS50</td>
<td>18871</td>
<td>0.031 0.027 0.025</td>
<td>0.015 0.013 0.014</td>
<td>0.93 0.94 0.93</td>
</tr>
<tr>
<td>met. tower</td>
<td>66777</td>
<td>0.049 0.047 0.044</td>
<td>0.025 0.023 0.023</td>
<td>0.90 0.89 0.89</td>
</tr>
<tr>
<td>$u_\star \theta_\star$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASFS30</td>
<td>22161</td>
<td>0.0091 0.0057 0.0045</td>
<td>0.0066 0.0043 0.0033</td>
<td>0.80 0.81 0.85</td>
</tr>
<tr>
<td>ASFS40</td>
<td>18498</td>
<td>0.0113 0.0071 0.0058</td>
<td>0.0089 0.0052 0.0041</td>
<td>0.70 0.74 0.78</td>
</tr>
<tr>
<td>ASFS50</td>
<td>18871</td>
<td>0.0056 0.0045 0.0049</td>
<td>0.0041 0.0033 0.0036</td>
<td>0.77 0.81 0.80</td>
</tr>
<tr>
<td>met. tower</td>
<td>66777</td>
<td>0.0073 0.0066 0.0062</td>
<td>0.0046 0.0040 0.0038</td>
<td>0.69 0.71 0.73</td>
</tr>
<tr>
<td>$u_\star q_\star$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASFS30</td>
<td>2692</td>
<td>3.20E-06 8.43E-07 1.06E-06</td>
<td>2.21E-06 5.83E-07 7.85E-07</td>
<td>0.75 0.63 0.63</td>
</tr>
<tr>
<td>ASFS50</td>
<td>3436</td>
<td>9.88E-07 8.93E-07 8.50E-07</td>
<td>6.55E-07 4.96E-07 4.63E-07</td>
<td>0.79 0.67 0.70</td>
</tr>
<tr>
<td>met. tower</td>
<td>33185</td>
<td>5.30E-07 5.94E-07 5.87E-07</td>
<td>2.97E-07 3.53E-07 3.43E-07</td>
<td>0.67 0.48 0.50</td>
</tr>
</tbody>
</table>
slightly lower RMSE (∼10%) when applied at 10 m compared with 2 m. This is as ex-pected: the 10 m gradients of the meteorological variables are larger than the correspond-ing 2 m gradients, so if the measurement errors at the different levels are similar in mag-nitude then the 10 m gradients should have lower relative error. Any inaccuracies in the estimated measurement heights should also be proportionally smaller at 10 m. Overall, the results are encouraging, with the ANN parametrizations consistently delivering per-formance improvements over the bulk algorithm.

Both methods produce similar estimates of the momentum flux $u_2^*$ and the resid-ual plots in Figure 1 share common features, such as a conservative bias (systematic underprediction of larger fluxes). However, the ANNs achieve a lower RMSE at all the MO-SAiC sites: a result which is robust under bootstrap resampling (Davison & Hinkley, 1997). The tendency of the ANNs to underpredict the magnitude of large and extreme fluxes was noted by Cummins et al. (2023) and is a known property of the models. In short, the ANNs have an inbuilt reluctance to extrapolate when faced with a combination of inputs not seen in training. That the bulk algorithm also underpredicts $u_2^*$ is unexpected and warrants investigation (see final paragraph of this section). Augmenting the ANN training set with data from MOSAiC further reduces the RMSE of the ANNs at all sites, as well as producing a visible attenuation of the conservative bias for larger fluxes. This result indicates that meteorological conditions conducive to large $u_2^*$ were consistent across the MOSAiC measurement sites.

The ANN parametrization strongly outperforms the bulk algorithm as an estima-tor of the sensible heat flux $u_\theta^*$, with RMSE 10-40 percent lower across the sites. It can be seen from Figure 2 that the improvements over the bulk are particularly apparent at the ASFS30 and ASFS40 stations. As was the case for $u_2^*$, the prediction errors of the two $u_\theta^*$ parametrizations share common features, including some clearly non-random deviations from the line $y = x$. In particular, there is a long tail in the panels for the met. tower in Figure 2. The tail is comprised of large negative fluxes whose magnitude was underestimated by both the bulk and ANN parametrizations. The underestimation occurred because, despite the large fluxes, the average temperature gradient over these 10-minute sampling periods was small. This highlights a limitation of the time-averaging approach to measuring fluxes and indicates that, potentially, new physics may be required to explain this phenomenon. It is possible that these large values in the direct measure-ments are the result of non-stationarity, i.e., that not all of the flux is really turbulence.
Figure 1. Predicted momentum fluxes $u_2^2$ at the MOSAiC sites in kinematic units plotted against observed eddy covariances. To the left are estimates obtained from a polar-specific bulk algorithm based on the Monin-Obukhov Similarity Theory; in the centre, estimates from the neural networks of Cummins et al. (2023); to the right, estimates from neural networks trained using MOSAiC-augmented data. The diagonal line $y = x$ would represent a perfect fit.
Figure 2. Predicted sensible heat fluxes $u, \theta$, at the MOSAiC sites in kinematic units plotted against observed eddy covariances. See Figure 1 caption for details.
At the tower, there is also evidence of an inverse phenomenon, where large gradients are observed but the corresponding fluxes are small. Including MOSAiC data from other sites in the ANN training set produces clear improvements at three of the four sites, with several systematic features in the residuals disappearing. The predictions at the ASFS50 site, however, became worse. Prediction errors at ASFS50 are the lowest for $u^2$ and $u_\theta^2$, so it doesn’t necessarily follow that the site is at fault. It is entirely possible that there are locally varying factors, not included in the set of input variables, which affect the fluxes (see final paragraph of this section). Identifying such variables has the potential to deliver further performance gains.

Latent heat flux $u_\ast q_\ast$ is by far the most difficult of the three fluxes to predict: $u_\ast q_\ast$ was generally small in magnitude at the MOSAiC measurement sites, suggesting a low signal-to-noise ratio. The ANNs are also disadvantaged here by a small training set from previous campaigns, comprised mainly of very small fluxes (Cummins et al., 2023). Results for $u_\ast q_\ast$ are therefore unsurprising: to the extent that the measured fluxes are small in magnitude, the ANNs perform well. For larger fluxes, the ANNs exhibit a strong conservative bias. Conversely, the bulk algorithm tends to overpredict the magnitude of $u_\ast q_\ast$.

It is because of these contrasting biases that the bulk achieves a higher correlation at the ASFS30 and ASFS50 stations, while at the same time the ANNs give RMSE reductions at those sites of about 70 and 10 percent respectively. At the MOSAiC tower, the bulk algorithm performs better, achieving a 10-percent lower RMSE. The ANNs trained on the MOSAiC-augmented data perform better at the ASFS50 site and the tower, but slightly worse at ASFS30. From Figure 3 it can be seen that the underestimation bias, while still present, is improved by training with MOSAiC data.

It should be noted that the biases, visible in the MOSAiC-augmented ANN predictions in Figures 1-3, should be further reduced by the next step, which is to incorporate all the MOSAiC data in the ANN training set. As more observations become available, we would expect the as-yet-unsampled regions of the input space to diminish, along with the associated biases. That is not to say that all biases can be resolved through more training data. As mentioned above, omission of important predictor variables has the potential to induce biases that will persist regardless of the volume of training data. For example, the upwards and downwards radiation terms are known to contribute significant explanatory power (e.g., Wulfmeyer et al., 2022). These radiation terms are nevertheless unsuitable for use as parametrization inputs, because the radiative fluxes in GCMs
Figure 3. Predicted latent heat fluxes $u \ast q \ast$ at the MOSAiC sites in kinematic units plotted against observed eddy covariances. See Figure 1 caption for details.
are themselves the output of complex parametrizations with their own errors and uncertainties. The surface characteristics may also be an important missing variable. In MOSAiC, the winter sea ice may have been generally aerodynamically rougher than that seen in the SHEBA campaign. This could potentially explain the underestimation of $u^2_\star$ by both the ANN and bulk algorithm parametrizations (see Figure 1).

5 Conclusions

Accurate representation in climate models of turbulent heat exchanges between the surface and atmosphere in polar regions is essential for constraining predictions of future climate change, locally and potentially globally. Surface turbulent fluxes in the polar boundary layer are currently parametrized using the traditional MOST, although alternative ML parametrizations based on ANNs have recently been proposed (Cummins et al., 2023). The wealth of new flux observations collected in the Arctic during the MOSAiC campaign has provided an excellent opportunity to validate and calibrate these alternative parametrizations.

In this study, the MOSAiC data have been used to validate ANN parametrizations of momentum, sensible heat and latent heat fluxes, that were originally trained on data from previous Arctic campaigns. The ANNs have been found to generalize well to the new data, particularly for momentum and sensible heat, yielding substantial reductions in error metrics such as RMSE when compared against a polar-specific bulk algorithm based on the MOST. Although the ANNs performed well at predicting small latent heat fluxes, limitations of the training data resulted in systematic underprediction of larger fluxes.

The ANN parametrizations, developed in Cummins et al. (2023) and validated in this study, have been recalibrated on an augmented training dataset that incorporates the observations from MOSAiC. These updated parametrizations have been implemented as a Fortran subroutine, suitable for deployment in climate models as a plug-in replacement for bulk algorithms (see Code Availability Statement). An important next step will be to perform sensitivity studies with these new parametrizations in a climate model.

In this way, the implications for the polar atmosphere and melting of Arctic sea ice can be assessed.
Code Availability Statement

Tools to reproduce the results presented in this study are publicly available at the following repositories:

- [https://doi.org/10.5281/zenodo.8207293](https://doi.org/10.5281/zenodo.8207293) This Python library provides functions to compute transfer coefficients and related variables (zeta, stability functions, aerodynamic and scalar roughness etc.), as well as to apply the bulk algorithm parametrizations used in this study.

- [https://doi.org/10.5281/zenodo.8207302](https://doi.org/10.5281/zenodo.8207302) This Python library provides functions to estimate meteorological parameters (humidity, latent heat as a function of temperature etc.).

- [https://doi.org/10.5281/zenodo.8207288](https://doi.org/10.5281/zenodo.8207288) This Fortran subroutine implements the neural network flux parametrizations developed in this study. The networks have been trained on all available datasets, including MOSAiC.

Data Availability Statement

MOSAiC campaign sites

The MOSAiC data used in this study are available from the National Science Foundation Arctic Data Center: met. tower ([https://doi.org/10.18739/A2PV6B3F](https://doi.org/10.18739/A2PV6B3F), Cox, Gallagher, Shupe, Persson, Blomquist, et al., 2023); ASFS30 ([https://doi.org/10.18739/A2FF3M18K](https://doi.org/10.18739/A2FF3M18K), Cox, Gallagher, Shupe, Persson, Grachev, et al., 2023a); ASFS40 ([https://doi.org/10.18739/A25X25F0P](https://doi.org/10.18739/A25X25F0P), Cox, Gallagher, Shupe, Persson, Grachev, et al., 2023b), ASFS50 ([https://doi.org/10.18739/A2XD0R00S](https://doi.org/10.18739/A2XD0R00S), Cox, Gallagher, Shupe, Persson, Grachev, et al., 2023c).

Pre-MOSAiC observational campaigns

The SHEBA data are available from the UCAR Earth Observing Laboratory: met. tower ([https://doi.org/10.5065/D65H7DNS](https://doi.org/10.5065/D65H7DNS), Andreas et al., 2007); PAM stations ([https://doi.org/10.5065/D62C8170](https://doi.org/10.5065/D62C8170), Andreas et al., 2012). The ACCACIA flight data are available from the CEDA archive: [https://doi.org/10.5285/0844186db1ba9e20319a2560f8d61651](https://doi.org/10.5285/0844186db1ba9e20319a2560f8d61651) (MASIN); [https://catalogue.ceda.ac.uk/uuid/c064b0c150274a1c9a8c563573f392e](https://catalogue.ceda.ac.uk/uuid/c064b0c150274a1c9a8c563573f392e) (FAAM). The ACSE cruise data are available from the CEDA archive ([https://doi.org/](https://doi.org/)}
10.5285/c6f1b1ff16f8407386e2d643bc5b916a, Brooks et al., 2022a). The AO16 cruise
data are available from the CEDA archive (https://doi.org/10.5285/614752d35dc147a598d5421443fb50e8,
Brooks et al., 2022b). The NSIDC sea ice concentration data are available from the NSIDC
archive (https://doi.org/10.7265/efmz-2t65, Meier et al., 2021).

Acknowledgments

This work was supported by a national funding by the Agence Nationale de la Recherche
within the framework of the Investissement d’Avenir program under the ANR-17-MPGA-
0003 reference. This article has received funding from the European Union’s Horizon 2020
research and innovation programme under grant agreement No 101003826 via project
CRiceS (Climate Relevant interactions and feedbacks: the key role of sea ice and Snow
in the polar and global climate system).

Data used in this manuscript were produced as part of the international Multidisci-
plinary drifting Observatory for the Study of Arctic Climate (MOSAiC) with the tag
MOSAiC20192020. We thank all persons involved in the expedition of the Research Ves-
sel Polarstern during MOSAiC in 2019-2020 (AWI_PS122_00) as listed in Nixdorf et al.
(2021). The MOSAiC observations received support from the US National Science Found-
ation Office of Polar Programs (OPP-1724551); the NOAA Physical Sciences Labora-
tory; and NOAA’s Global Ocean Monitoring and Observing Program (FundRef https://
doi.org/10.13039/100018302).

This is a contribution to the Year of Polar Prediction (YOPP), a flagship activ-
ity of the Polar Prediction Project (PPP), initiated by the World Weather Research Pro-
gramme (WWRP) of the World Meteorological Organisation (WMO). We acknowledge
the WMO WWRP for its role in coordinating this international research activity.

Neural network ensembles were fitted using the R package caret (Kuhn & John-
son, 2013), which itself depends on the R package nnet (Venables & Ripley, 2002) to fit
the underlying models. Bootstrapping of model performance metrics was performed us-
ing the R package boot (Canty & Ripley, 2022).

References


change over the marginal ice zone and recommendations for its parametrization. 

Atmospheric Chemistry and Physics, 16(3), 1545–1563. doi: 10.5194/acp-16-1545-2016


