White Rose Research Online
eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/228114/

Version: Accepted Version

Article:

Cummins, D.P., Guemas, V., Blein, S. et al. (2024) Reducing Parametrization Errors for Polar Surface Turbulent Fluxes Using Machine Learning. Boundary-Layer Meteorology, 190. 13. ISSN: 0006-8314

https://doi.org/10.1007/s10546-023-00852-8

© The Author(s), under exclusive licence to Springer Nature B.V. 2024. This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature’s AM terms of use (https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms), but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/s10546-023-00852-8.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Reducing parametrization errors for polar surface turbulent fluxes using machine learning

Donald P. Cummins[1], Virginie Guemas[1], Sébastien Blein[1], Ian M. Brooks[2], Ian A. Renfrew[3], Andrew D. Elvidge[3], and John Prytherch[4]

[1]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France
[2]School of Earth & Environment, University of Leeds, Leeds, LS2 9JT, UK
[3]School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, UK
[4]Department of Meteorology, Stockholm University, Stockholm, Sweden

September 6, 2023

Abstract

Turbulent exchanges between sea ice and the atmosphere are known to influence the melting rate of sea ice, the development of atmospheric circulation anomalies and, potentially, teleconnections between polar and non-polar regions. Large model errors remain in the parametrization of turbulent heat fluxes over sea ice in climate models, resulting in significant uncertainties in projections of future climate. Fluxes are typically calculated using bulk formulae, based on Monin-Obukhov similarity theory, which have shown particular limitations in polar regions. Parametrizations developed specifically for polar conditions (e.g. representing form drag from ridges or melt ponds on sea ice) rely on sparse observations and thus may not be universally applicable.

In this study, new data-driven parametrizations have been developed for surface turbulent fluxes of momentum, sensible heat and latent heat in the Arctic. Machine learning has already been used outside the polar regions to provide accurate and computationally inexpensive estimates of surface turbulent fluxes. To investigate the feasibility of this approach in the Arctic, we have fitted neural-network models to a reference dataset (SHEBA). Predictive performance has been tested using data from other observational campaigns. For momentum and sensible heat, performance of the neural networks is found to be comparable to, and in some cases substantially better than, that of a state-of-the-art bulk formulation.

These results offer an efficient alternative to the traditional bulk approach in cases where the latter fails, and can serve to inform further physically based developments.

1 Introduction

Arctic sea ice has declined drastically over recent decades, drawing special attention and instigating concern about potential repercussions on local ecosystems (Kovacs et al., 2011; Post et al., 2013; Tynan, 2015), indigenous populations (Meier et al., 2014) and lower-latitude climate (Cohen et al., 2014; Jung et al., 2015; Cohen et al., 2020; Liu et al., 2022). Indeed, since the advent of satellite imagery, about three million km^2 of Arctic sea ice has been lost; this represents a decrease by half of sea ice extent at the end of the summer season and has coincided with a reduction in ice thickness also by about a half (Gascard et al., 2019). Arctic sea ice loss is expected to continue in coming decades, although large uncertainties still remain regarding the expected rate of disappearance (e.g. Bonan et al., 2021a,b). Surface warming two to four times faster than the global mean has also been observed in the Arctic over the last couple of decades (e.g. Cohen et al., 2014). This phenomenon is commonly referred to as Arctic amplification (e.g. Serreze and Francis, 2006; Graversen et al., 2008; Serreze and Barry, 2011). Heat exchanges between the ice and the atmosphere are a key determinant of both the rate of Arctic sea ice melting (e.g. Rothrock et al., 1999; Screen and Simmonds, 2010) and the Arctic amplification (e.g.

Serreze et al., 2009; Lesins et al., 2012; Previdi et al., 2021).

Large uncertainties still remain in the estimation of surface turbulent fluxes under polar-specific conditions (e.g. Vihma et al., 2014), where the atmospheric boundary layer is frequently stable and turbulence can be intermittent (e.g. Andreas, 1998). Turbulent fluxes are typically modelled through bulk formulae based on the Monin-Obukhov similarity theory (Monin and Obukhov, 1954; Garratt, 1994; Andreas, 1998). Stability corrections tailored to polar conditions have been proposed (Grachev et al., 2007) from the year-long SHEBA (Surface Heat Budget of the Arctic Ocean, Uttal et al., 2002; Persson et al., 2002) campaign, but the number of measurements used to produce them is still limited compared to the volume of data available from tropical areas used to tailor stability corrections for convective conditions. Polar-specific parametrizations have also been proposed for surface roughness (e.g. Andreas, 1987; Andreas et al., 2010b; Andreas, 2011) and for the form drag arising from alternating floes and leads (e.g. Lüpkes et al., 2012; Lüpkes and Gryanik, 2015; Elvidge et al., 2016). However, further calibration of those parametrizations might still be necessary with upcoming campaigns (e.g. Elvidge et al., 2021). Such parametrizations have been shown to reduce polar biases in atmospheric models (Renfrew et al., 2019; Elvidge et al., 2023). Most climate models still use a constant and fixed neutral transfer coefficient for all ice types and thicknesses (e.g. Notz et al., 2013; Lüpkes et al., 2013). Reluctance to include the most up-to-date polar-specific parametrizations in climate models has originated partly from the lack of observational data available to validate and calibrate those parametrizations, and from inadequacies in model formulation. This situation has improved with the recent Year of Polar Prediction (Jung et al., 2016) and the MOSAiC campaign (Shupe et al., 2022), as well as other studies using data gathered from recent intensive field campaigns in the Arctic (Elvidge et al., 2016, 2021; Srivastava et al., 2022). Another obstacle to widespread adoption of advanced schemes for specific situations (such as the presence of sea ice) may be the fact that large-scale models have historically struggled to represent the stable boundary layer, even in simple, homogeneous conditions (e.g. the GABLS experiments, Cuxart et al., 2006; Svensson et al., 2011; Bosveld et al., 2014), although recent work suggests that better parameter calibration is needed (Audouin et al., 2021). Finally, both the stability correction and the surface roughness parametrizations, which are essential components of the Monin-Obukhov similarity theory leading to estimates of surface turbulent fluxes, depend on the surface turbulent fluxes themselves. Hence, iterative algorithms are often used (Fairall et al., 2003; Edson et al., 2004). Calibration of such algorithms on noisy data is complicated by the “self-correlation” phenomenon (e.g. Baas et al., 2006).

Some of those challenges could be bypassed through the use of machine-learning algorithms, a type of empirical “black box” relating the turbulent fluxes directly to the observable meteorological quantities on which they depend. Although somewhat lacking in physical interpretability, the data-driven approach avoids the many sequential steps involved in developing bulk schemes, which are likely to result in compounding errors. There is also the possibility of assessing the effect of input variables not directly included in the physical relationships of the bulk method. Most importantly, machine-learning models have the potential to deliver more accurate flux estimates. The observational data collected thus far offer the opportunity to assess the relevance of such an approach. Machine learning has already been tested for the estimation of surface turbulent fluxes outside the polar regions with a successful outcome (e.g. Pelliccioni et al., 1999; Qin et al., 2005a,b; Wang et al., 2017; Safa et al., 2018; Xu et al., 2018; Leufen and Schädler, 2019; Wang et al., 2021). In recent studies, machine-learning parametrizations have been developed which can accurately estimate turbulent fluxes observed at measurement towers (McCandless et al., 2022; Wulfmeyer et al., 2022), and the properties of such parametrizations investigated in large-eddy simulations (Muñoz-Esparza et al., 2022). In this study, we apply data-driven methods to the SHEBA campaign to investigate the viability of this approach in polar conditions, and we assess its performance compared to a state-of-the-art bulk formulation over a few other recent Arctic campaigns. This article is structured as follows. Section 2 describes the available field data used to carry out our analysis. Section 3 presents the bulk- and the data-driven approaches to estimate surface turbulent fluxes as well as the statistical methodology to assess their performance. Section 4 describes the results. Conclusions and perspectives are provided in Section 5.

2 Data

Our analyses make use of data from the following observational campaigns.

- SHEBA (Surface Heat Budget of the Arctic Ocean): the SHEBA ice camp was set up around the icebreaker *Des Groseilliers*, which was frozen into the Arctic ice pack and drifted approximately 2700 km in the Beaufort Gyre between 2 October 1997 and 11 October 1998. It started in the Beaufort Sea, drifted westward into the Chukchi Sea, then turned north into the Arctic Ocean near the date line. Turbulent fluxes and mean meteorological data were measured continuously at five levels: 2.2, 3.2, 5.1, 8.9, and 18.2 m (or 14 m during most of the winter), on the 20-metre main SHEBA tower. Turbulent covariances and variances were estimated at each level based on one-hour averaging and derived through the frequency integration of the cospectra and spectra (Persson et al., 2002). From the more than 8000 h for which the main SHEBA tower was instrumented, over 6000 h passed quality control. Four remote sites, ranging in distance from 0.25 to 30 km from the main camp and known as Portable Automated Mesonet (PAM) stations, were also instrumented. The PAM stations provided measures of wind, temperature and humidity together with estimates of surface heat fluxes through eddy covariances. The sea ice characteristics changed radically during the year-long deployment, from compact and snow-covered in winter (Andreas et al., 2010b), through to a covering littered with deep melt ponds and leads in summer (Andreas et al., 2010a). Other details of the SHEBA programme, the ice camp, deployed instruments, data processing, accuracy, calibration, and archived data files may be found in Andreas et al. (1999, 2002, 2003, 2006); Persson et al. (2002); Uttal et al. (2002); Grachev et al. (2002, 2005).

- ACCACIA (Aerosol-Cloud Coupling and Climate Interactions in the Arctic): eight flights, from 21 to 31 March 2013, to the northwest of Svalbard over the Fram Strait and to the southeast of Svalbard in the Barents Sea, were conducted with two aircraft: a DHC6 Twin Otter operated by the British Antarctic Survey and equipped with the Meteorological Airborne Science Instrumentation (MASIN) (King et al., 2008; Fiedler et al., 2010); and the UK Facility for Airborne Atmospheric Measurement (FAAM) BAe-146 (Renfrew et al., 2008; Petersen and Renfrew, 2009). Both aircraft measured turbulent fluxes and meteorological parameters in the atmospheric boundary layer, as well as surface meteorological parameters through radar, leading to more than 200 new estimates of surface drag in the Marginal Ice Zone (Elvidge et al., 2016). In the Barents Sea, sea ice was characterized by small, unconsolidated ice floes (generally associated with a higher neutral drag coefficient), while over the Fram Strait there were typically larger, smoother floes (Elvidge et al., 2016).

- ACSE (Arctic Cloud in Summer Experiment): the icebreaker *Oden* left Tromsö, Norway, on 5 July 2014, crossing the Kara, Laptev, East Siberian, and Chukchi Seas, following the Siberian Shelf, and arriving in Barrow, Alaska, on 18 August. A second leg left Barrow on 21 August following a similar route back, albeit farther north. The expedition ended on 5 October in Tromsö. An instrumented mast at the bow of the ship was used to obtain surface turbulent fluxes through the eddy covariance technique (Prytherch et al., 2017; Thornton et al., 2020; Srivastava et al., 2022), while a weather station on the seventh deck at about 25 m measured temperature, humidity and wind (Sotiropoulou et al., 2016). The 12-week ACSE cruise took place within the Arctic pack ice, working over, and just off the Siberian Shelf, during summer melt and early autumn freeze-up conditions. It sampled a wide range of sea ice morphologies (Persson et al., 2015; Srivastava et al., 2022).

- AO16 (Arctic Ocean 2016): the icebreaker *Oden* departed from Longyearbyen, Svalbard, on 8 August and operated in the Arctic Ocean, mainly in the Amundsen Basin and in areas around the underwater mountain ranges, Lomonosov Ridge and Alpha Ridge, until 19 September 2016. Similar instrumentation as in the ACSE expedition was employed (Tjernström and Jakobsson, 2021; Srivastava et al., 2022). The six-week AO16 cruise followed a more northerly route than ACSE (mostly north of 85°N) and the surface was mostly characterized by old and thick ice, with intermittent patches of thin ice and melt ponds (Srivastava et al., 2022).

Corresponding estimates of sea ice concentration from satellite imagery were obtained from the National Snow and Ice Data Center (NSIDC, Meier et al., 2021). For the AO16 dataset, the included satellite-based estimates were used. See the data availability statement for details of where to obtain the datasets used in this study.

3 Methods

3.1 Bulk flux parametrizations

Sea ice, atmosphere and coupled climate models rely on Monin-Obukhov Similarity Theory (MOST, Monin and Obukhov, 1954) to represent turbulent fluxes at the surface-atmosphere interface through bulk flux parametrizations. The bulk approach consists in estimating surface turbulent fluxes of momentum τ, sensible heat H_S and latent heat H_L from the near-surface gradient of model-resolved (or averaged) variables (wind speed, temperature and humidity), weighting each gradient by the corresponding transfer coefficient for momentum (C_D, the drag coefficient), sensible heat (C_H) and latent heat (C_q):

$$
\begin{aligned}
|\tau| &= \rho C_D(z) u(z)^2, && (1)\\
H_S &= \rho c_p C_H(z) u(z) \Big(\theta_s - \theta(z)\Big), && (2)\\
H_L &= \rho L_v C_q(z) u(z) \Big(q_s - q(z)\Big), && (3)
\end{aligned}
$$

where ρ is the air density, u is the horizontal component of wind, θ is the potential temperature, q is the specific humidity, c_p is the specific heat of air at constant pressure, L_v is the latent heat of vaporization or sublimation, C_D, C_H and C_q are the transfer coefficients which all depend on the height z at which the wind speed u, the temperature θ and the humidity q are taken, θ_s is the surface potential temperature and q_s is the surface humidity.

A bulk parametrization essentially consists of an algorithm estimating the transfer coefficients C_D, C_H and C_q. In this study, we use a composite bulk algorithm comprising recent components developed for use in polar conditions. Over the ocean, we use the COARE 3.0 algorithm (Fairall et al., 2003; Edson et al., 2013), with a first guess of transfer coefficients deduced from the stability according to Grachev and Fairall (1997) to speed up the convergence, stability correction for unstable conditions from Grachev et al. (2000), and for stable conditions from Beljaars and Holtslag (1991), the aerodynamic roughness model from Edson et al. (2013) and the scalar roughness model from Fairall et al. (2003). The COARE 3.0 algorithm has been extensively tested and is currently used in large-scale climate models (e.g. CNRM-CM6, Voldoire et al., 2019).

Over sea ice, the stability correction under stable conditions relies on Grachev et al. (2007), the scalar roughness model on Andreas (1987) and the aerodynamic roughness model on Andreas et al. (2010b), the rest of the algorithm being the same as over the ocean (i.e. COARE 3.0). Sea ice concentrations between zero and unity are handled by taking a weighted average of the estimated fluxes over ocean and sea ice, weighting by sea ice concentration. This averaging approach, known as the *mosaic* method (e.g. Vihma, 1995), is already used in general circulation models (GCMs), such as CNRM-CM6. For the momentum flux, we also include a form drag contribution, which accounts for the increased turbulence observed where floes and leads are alternating (Lüpkes and Gryanik, 2015). This combination leads to the best estimates of surface turbulent fluxes that we can obtain in light of the most recent studies focusing on polar regions (Andreas et al., 2010a,b; Lüpkes and Gryanik, 2015; Elvidge et al., 2016; Srivastava et al., 2022). On the datasets used in this study, momentum flux estimates from our polar-specific algorithm have up to 23 % lower root-mean-square error (RMSE) than those from the unmodified COARE 3.0. Estimated heat fluxes are very similar with or without the polar-specific modifications (corr. > 0.99). The bulk algorithm described here is publicly available for download as a Python library (see the code availability statement for a link to the repository) and the equations of the polar-specific components are given in the appendix.

Following the MOST, surface turbulent fluxes are commonly expressed as functions of the scaling parameters $u_\star$, $\theta_\star$ and $q_\star$ for wind, potential temperature and humidity respectively:

$$|\tau| = \rho u_\star^2, \tag{4}$$

$$H_S = -\rho c_p u_\star \theta_\star, \tag{5}$$

$$H_L = -\rho L_v u_\star q_\star. \tag{6}$$

So-called kinematic fluxes $u_\star^2$, $u_\star\theta_\star$ and $u_\star q_\star$ will be used in the remainder of this study because they correspond directly to the eddy covariances which are measured in the field.

3.2 Neural networks

Multiple machine-learning algorithms were tested for this study including random forests (Breiman, 2001), gradient boosting machines (Friedman, 2001; Chen and Guestrin, 2016), generalized additive models (Hastie and Tibshirani, 1986) and multivariate adaptive regression splines (Friedman, 1991). In general, it was found that algorithms permitting high-degree interactions between input variables performed better, but often at the cost of many free parameters. The algorithm chosen for the final analysis was the artificial neural network, which we found to offer the best predictive performance and with a relatively parsimonious model.

Neural networks can approximate continuous functions of multiple variables (Hornik et al., 1989) and have performed well in previous studies as estimators of surface turbulent fluxes (Pelliccioni et al., 1999; Qin et al., 2005a; Wang et al., 2017; Safa et al., 2018; Xu et al., 2018; Leufen and Schädler, 2019; McCandless et al., 2022; Muñoz-Esparza et al., 2022; Wulfmeyer et al., 2022). Many specialized configurations of network nodes or *architectures* have been developed for specific applications, for example the convolutional networks used for image recognition (e.g. Krizhevsky et al., 2017), however in this study attention is restricted to the simplest architecture: the single-layer feed-forward network. At its most basic, a neural network is a non-linear generalization of linear regression, in which the independent and dependent variables are represented as input and output nodes. A feed-forward network passes information from a layer of input nodes, by way of linear combination, to a so-called hidden layer, containing nodes with non-linear activation functions (see Figure 1). An activation function is simply a non-linear transformation. Finally, the output of the hidden layer is combined in the output layer, which in our case consists of a single node representing the estimated flux. For an introduction to feed-forward neural networks see Ripley (1996).

Let x_i denote a seven-element input vector containing measurement height z, wind speed $u(z)$, potential temperature $\theta(z)$, surface temperature θ_s, specific humidity $q(z)$, surface humidity q_s and sea ice concentration C_i. These are the predictor variables used as inputs to the bulk algorithm described above. Letting $y \in \{u_\star^2, u_\star\theta_\star, u_\star q_\star\}$ denote the flux of interest in kinematic units, we can write a network-predicted flux $\hat{y}$ as

$$\hat{y} = \alpha_0 + \sum_{j=1}^{N_j} w_j f\left(\alpha_j + \sum_{i=1}^{7} w_{ji} x_i\right), \tag{7}$$

where, in neural-network parlance, constants α_j are known as biases, coefficients w_{ji} are weights, and f is a non-linear activation function. Here we use the sigmoid activation $f(x) = e^x/(1+e^x)$. The number of hidden nodes N_j controls the total number of parameters and hence the complexity of the network. In practice, it is necessary to rescale the inputs and outputs, to avoid excessive saturation of each hidden node's activation function (Ripley, 1996). Saturation refers here to the disappearing gradient of the sigmoid function for inputs with large magnitude. We rescale all variables for model training via z-score normalization.

Training of the neural networks was performed by minimizing the sum of squared prediction errors on a designated training set. The sum of squared errors was minimized using the implementation of the Broyden-Fletcher-Goldfarb-Shanno (BFGS, Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) quasi-Newton method in the *nnet* package for R (Venables and Ripley, 2002; R Core Team, 2021). Numerical optimization of the network parameters requires a non-zero initialization of the weights and biases, which is obtained by random number generation. Due to the presence of local minima in the least-squares objective function to be minimized, such random initialization introduces randomness in the fitted networks, making it difficult to measure predictive performance.

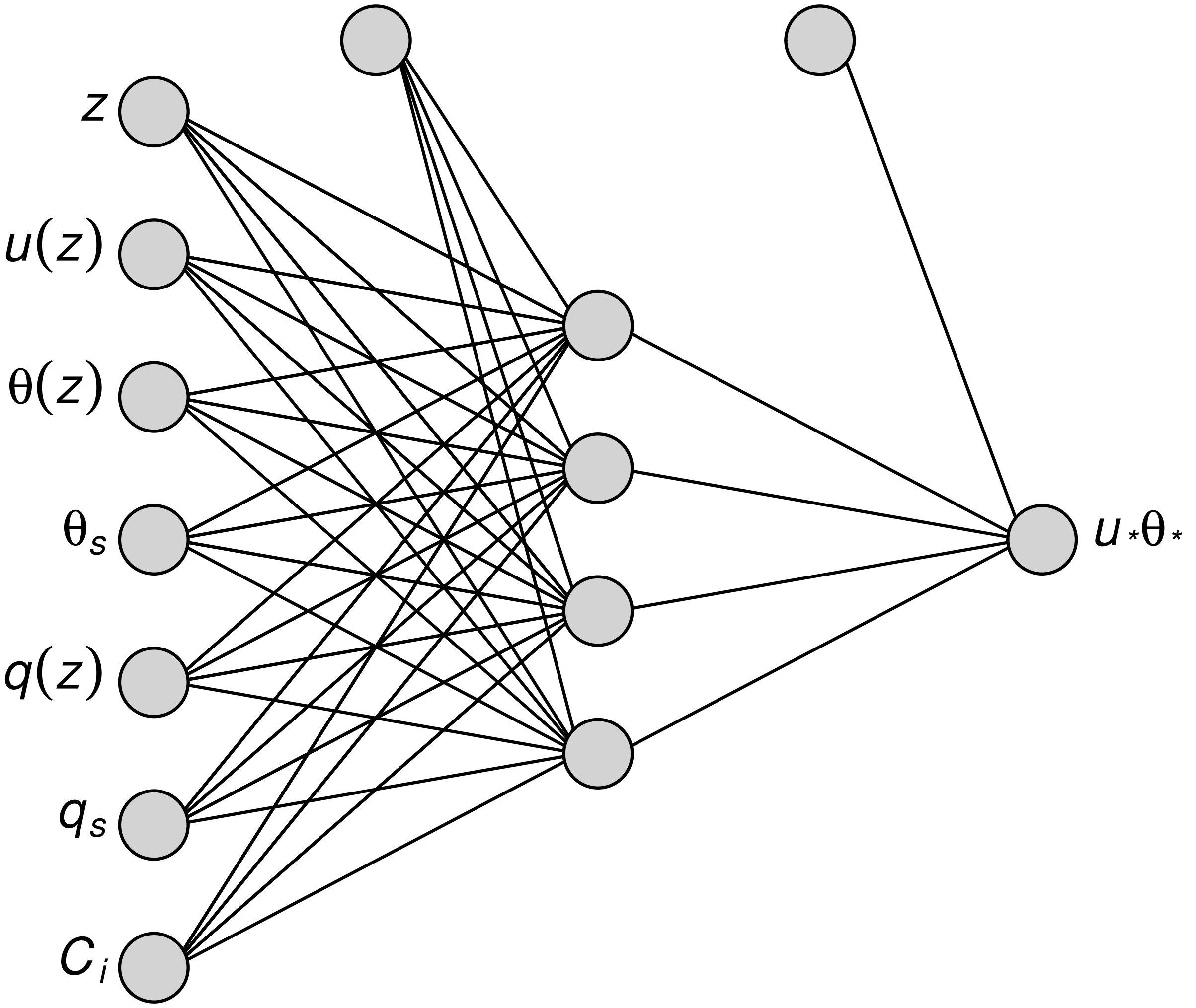

Figure 1: Neural-network architecture schematic. Information is passed from the seven inputs nodes on the left, through the four hidden nodes with non-linear activation functions, to the output node - in this case kinematic sensible heat flux $u_\star\theta_\star$. Separate networks with the same architecture were estimated for fluxes of momentum $u_\star^2$ and latent heat $u_\star q_\star$. The two unlabelled nodes at the top correspond to node biases (analogous to the intercept term in a linear regression) and represent constant inputs of value unity.

Following Ripley (1996) we take two steps to address this problem. Firstly, applying a quadratic penalty to the network weights (a technique known as weight decay, L2 regularization or Tikhonov regularization) constrains the weights to be small in magnitude and increases the convexity of the objective, thus improving the numerical conditioning of the minimization problem. Secondly, we fit each neural network 100 times, each time using an independent random parameter initialization. The set of 100 fitted networks is used as a model ensemble, with an overall prediction obtained by averaging predictions from all the members. Thus we largely eliminate randomness in our results due to parameter initialization, and at the same time we gain a small boost in mean predictive performance compared with using single networks. The performance gain from ensemble averaging is sufficiently small (< 5 % reduction in RMSE) that, for implementation in a GCM, the improvement over single networks may not be worth the computational cost.

The neural networks in this study have two tunable hyperparameters: N_j, the number of nodes in the hidden layer; and λ, the L2 regularization penalty. Setting $\lambda = 0.01$, at the upper bound of the range recommended by Ripley (1996), and $N_j = 4$ gives the networks more than enough flexibility to emulate their respective bulk-algorithm counterparts over the range of meteorological conditions observed. Since the bulk algorithm represents an *a priori* estimate of the functions to be approximated by the network, it makes sense to use the bulk in this way when choosing the model complexity. Aggressive tuning of hyperparameters using observational data has been avoided, so as not to excessively bias the cross-validation results (see Section 3.3), although it should be noted that these settings are roughly optimal. Increasing N_j beyond four does not substantially improve performance for any of the fluxes considered, while the chosen value of λ is sufficient to keep network weights within a suitable range for the sigmoid activation (standard deviation of order unity).

Our approach to neural-network modelling of surface turbulent fluxes differs in some ways from recent studies by other authors. For example, McCandless et al. (2022), Muñoz-Esparza et al. (2022) and Wulfmeyer et al. (2022) employ deep networks, i.e. networks containing multiple hidden layers, as well as more nodes per layer, resulting in models with orders of magnitude more parameters. The rationale behind that approach is to equip the network with a huge learning capacity (flexibility to approximate functions), thus ruling out any possibility of underfitting. To mitigate the corresponding tendency of larger models to overfit, those studies used a reduced number of training iterations. The network architecture used in this study is more similar to that of Leufen and Schädler (2019), who found that single-layer networks with just a handful of nodes provide a good balance of parsimony and flexibility for modelling turbulent fluxes. Indeed, the purpose of regularization techniques, such as early stopping during training, is to reduce the effective degrees of freedom of the model (Hastie et al., 2009). Since the ultimate goal is to obtain parametrizations suitable for use in a GCM, where redundant computation should be avoided wherever possible, we therefore favour smaller, more highly optimized networks. The neural-network parametrization of Leufen and Schädler (2019) had roughly the same computational cost as a bulk algorithm based on the MOST, when implemented in a one-dimensional land surface model.

Our model also differs from those in the above studies in our choice of input variables to the network. Specifically, network inputs in the present study were constrained to be the same variables used as inputs to the bulk algorithm. This restriction allows a fair apples-to-apples comparison of the performance of the different parametrizations. Wulfmeyer et al. (2022) use incoming and outgoing radiation as predictor variables. While we second their finding that the radiation terms contribute explanatory power to predict fluxes, we have refrained from including them in our parametrizations. This is because the surface radiation balance in a GCM is itself dependent on complicated radiative transfer parametrizations, which means that including radiation terms in a turbulent flux model risks introducing another layer of compounding errors. We have also refrained from pre-computing gradients of temperature and humidity, or other derived quantities such as the bulk Richardson number. Instead, by including measurement height z as a model input, we allow any explicit height dependence of the fluxes to be modelled empirically. Thus we avoid interpolating observations to reference heights, which would anyhow be impossible (given that, except at the SHEBA tower, most of the variables were measured at only one height above ground level) without making parametric assumptions on the vertical profiles (e.g. the MOST). From the ACCACIA flight data, only those fluxes observed at altitudes below 30 m have been used in model training and validation, due to concern about how representative the higher-altitude observations (up to 90 m) are of surface-layer conditions. The first model level in GCMs, e.g. CNRM-CM6, is typically well below 30 m over ocean.

3.3 Cross validation

The purpose of the present study is to determine whether a machine-learning algorithm, trained on observational data collected in the Arctic, can offer improved performance over advanced bulk formula parametrizations developed for use in polar regions. An important property of neural-network models is that they can often achieve an arbitrarily high goodness-of-fit on the training data, simply by increasing the number of free parameters in the model (Hastie et al., 2009). Therefore, when observations contain measurement errors, as is usually the case, one must be careful not to overfit the data (extract not only signal but also noise). To mitigate the risk of overfitting, the performance of a machine-learning model should be assessed based on its ability to predict *unseen* data (data not used to train the model), hereafter the "out-of-sample" performance. A popular approach to measuring out-of-sample performance is K-fold cross validation, which entails splitting the data into K independent subsets. For each $k \in \{1, \ldots, K\}$, the model is fitted to those data *not* contained in subset k. The fitted model is then used to predict the response variable in subset k. Iterating over the K subsets, one thus obtains a complete dataset of out-of-sample predictions from which to compute performance metrics such as root-mean-square error (RMSE) or mean absolute error (MAE).

Statistical independence of the K subsets is a necessary condition for cross-validated performance metrics to be meaningful. In a typical machine-learning problem, genuinely independent subsets are not available and are instead manufactured by randomly partitioning a single dataset into K folds. Datasets of surface turbulent fluxes are typically measured as time series, in our case none longer than a year, which invalidates the independence assumption of the random subsetting approach. While one alternative would be to construct an elaborate stratified cross-validation scheme, intended to account for temporal (and other sources of) autocorrelation, we instead opt to make use of the fact that in this study we have a database of observations from multiple field campaigns. The observations were collected at a range of locations, at different times, using different measurement instruments, by different teams of researchers and under different meteorological conditions. For example, data were collected at towers, from ship masts and from low-flying aircraft. In theory, cross validation of a flux parametrization across observational campaigns estimates how well we would expect it to perform at predicting data collected in a future campaign, which would seem an appealingly objective metric for comparison. In practice, however, there are some limitations on the independence of the available datasets. For example, overlap in the large-scale meteorological conditions at the SHEBA tower and SHEBA PAM stations cannot be ruled out as a potential source of statistical dependence. Nevertheless, the degree of heterogeneity in the present ensemble of datasets is such that achieving a generalizable parametrization may be regarded as a real benchmark of success.

3.4 Variable importance

One advantage of supplying the same input variables to the neural networks as to the bulk algorithm is that the two methods can be diagnosed and compared using *variable importance* techniques. Wulfmeyer et al. (2022) used the so-called “feature importance” methodology proposed in Breiman (2001) to compare two machine-learning algorithms for computing surface turbulent fluxes. The feature importance algorithm is as follows. Firstly, a model is used to make predictions on a dataset and the quality of those predictions is assessed using some error metric (e.g. MSE). Then, for each input variable to the model x_i, the values of that variable are randomly shuffled in the dataset and the model predictions (and corresponding error metrics) recomputed. The importance of each input variable is estimated as the respective amount by which the performance metric increases when that variable is shuffled. This procedure has been applied to the

Table 1: Cross-validated performance metrics of neural-network (nnet) and bulk-algorithm flux parametrizations in kinematic units. The cross validation was over observational campaigns. Boldface indicates a significantly better score at the five-percent level in one of root-mean-square error (RMSE), mean absolute error (MAE) or Pearson correlation.

			RMSE		MAE		corr.	
	campaign	n	bulk	nnet	bulk	nnet	bulk	nnet
$u_\star^2$	accacia	65	**0.085**	0.247	**0.047**	0.234	0.57	0.49
	acse	2324	0.053	0.051	0.036	0.035	**0.83**	0.80
	ao16	250	0.090	**0.072**	0.063	**0.044**	0.82	**0.90**
	sheba_atlanta	3203	0.019	0.019	0.012	**0.011**	0.91	**0.91**
	sheba_baltimore	279	**0.036**	0.038	**0.021**	0.022	0.64	**0.65**
	sheba_csm	225	0.029	0.029	0.019	0.019	0.89	0.88
	sheba_florida	427	0.048	**0.047**	0.031	**0.031**	0.82	**0.82**
	sheba_tower	23028	**0.018**	0.021	**0.011**	0.014	**0.95**	0.93
$u_\star\theta_\star$	accacia	50	**0.0214**	0.0476	**0.0163**	0.0273	**0.95**	0.61
	acse	796	**0.0095**	0.0115	**0.0064**	0.0083	**0.85**	0.81
	ao16	95	0.0154	**0.0135**	0.0102	**0.0093**	0.58	**0.73**
	sheba_atlanta	3203	0.0063	**0.0055**	0.0040	**0.0035**	0.65	0.64
	sheba_baltimore	279	0.0061	0.0062	**0.0043**	0.0050	0.48	0.45
	sheba_csm	225	0.0202	**0.0177**	0.0093	**0.0085**	0.00	0.01
	sheba_florida	427	0.0141	**0.0132**	0.0102	**0.0093**	0.29	**0.32**
	sheba_tower	23633	**0.0037**	0.0061	**0.0025**	0.0046	**0.82**	0.47
$u_\star q_\star$	accacia	34	**2.85E-06**	7.96E-06	**1.76E-06**	6.24E-06	**0.97**	0.72
	acse	673	1.03E-05	1.05E-05	5.93E-06	6.24E-06	**0.35**	0.17
	ao16	81	1.44E-05	1.46E-05	6.27E-06	**5.58E-06**	0.24	0.10
	sheba_tower	13511	**1.06E-06**	5.72E-06	**5.51E-07**	4.03E-06	**0.69**	-0.23

bulk algorithm and neural-network models used in this study, and the resulting variable importances scaled to lie between zero and unity.

4 Results

Performance metrics computed from the cross-validation experiments are given in Table 1. Measured differences in performance between the bulk and neural-network methods were tested for statistical significance on a per-dataset basis using a bootstrapping approach (Davison and Hinkley, 1997). Model-predicted turbulent fluxes in kinematic units are plotted against field observations from the ACSE dataset in Figure 2. Results for

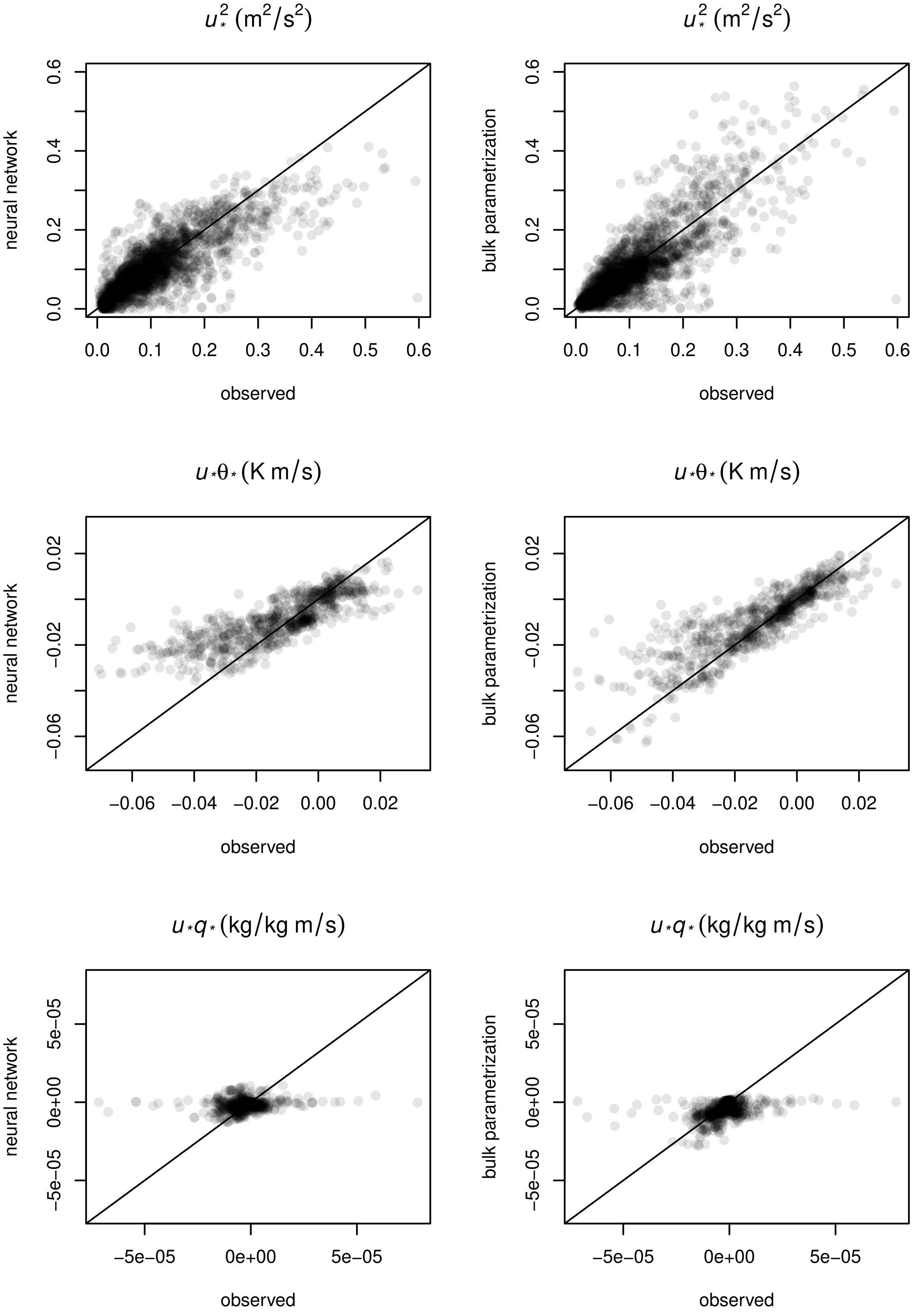

Figure 2: Scatter plots of predicted vs observed fluxes in kinematic units for the ACSE campaign. The diagonal line has equation $y = x$ and represents 100 % predictive accuracy.

ACSE are fairly representative of the results across the other datasets, with the notable exception of the ACCACIA flight data. Equivalent figures for the other datasets, including ACCACIA, are provided in the supplementary information. Computed variable importance metrics for the bulk algorithm and neural-network models are shown in Figure 3.

It can be seen from these results that, of the three turbulent fluxes $\{u_\star^2, u_\star\theta_\star, u_\star q_\star\}$, the momentum flux $u_\star^2$ is by far the easiest to predict, due to its strong linear correlation with squared wind speed u^2. Compared with the bulk algorithm, the neural-network model performs similarly well or better on most datasets, except for ACCACIA and the SHEBA tower. Performance at the SHEBA tower is expected to be worse because that dataset accounts for most of the available observations and removing it reduces the training set to a small fraction of its former size. In ACCACIA, the neural-network model systematically overpredicts $u_\star^2$, in an unsuccessful attempt to extrapolate beyond the range of measurement heights z seen in the training set. Note that z in ACCACIA ranges from 20 to 30 m, while the highest z in the other datasets is $z = 20.3$ m in ACSE and AO16. The greatest proportional error reduction was seen in the AO16 dataset where the neural-network model gave a 21 % reduction in RMSE. In general, bulk estimates of $u_\star^2$ are very good, so there is little room for large performance improvements, especially considering the fact that we are dealing with noisy real-world observations. There is a notable negative bias in the neural-network estimates for large values of $u_\star^2$ in the ACSE dataset (see Figure 2), which is an artifact of the conservative extrapolatory properties of neural networks (see next paragraph). It should be noted that the bulk algorithm exhibits a similar negative bias in other datasets, e.g. AO16. Due to the lack of boundary constraints on its output, the neural-network model occasionally produces very small negative estimates of $u_\star^2$. In this study, such estimates have been rounded up to zero before computing performance metrics, as this step can easily be included when the parametrization is implemented in a GCM. Estimation of $u_\star^2$ in a log-transformed space, a common method for enforcing strict positivity, is inappropriate here, since in a GCM it is the magnitude of the flux in an absolute rather than relative sense which is important.

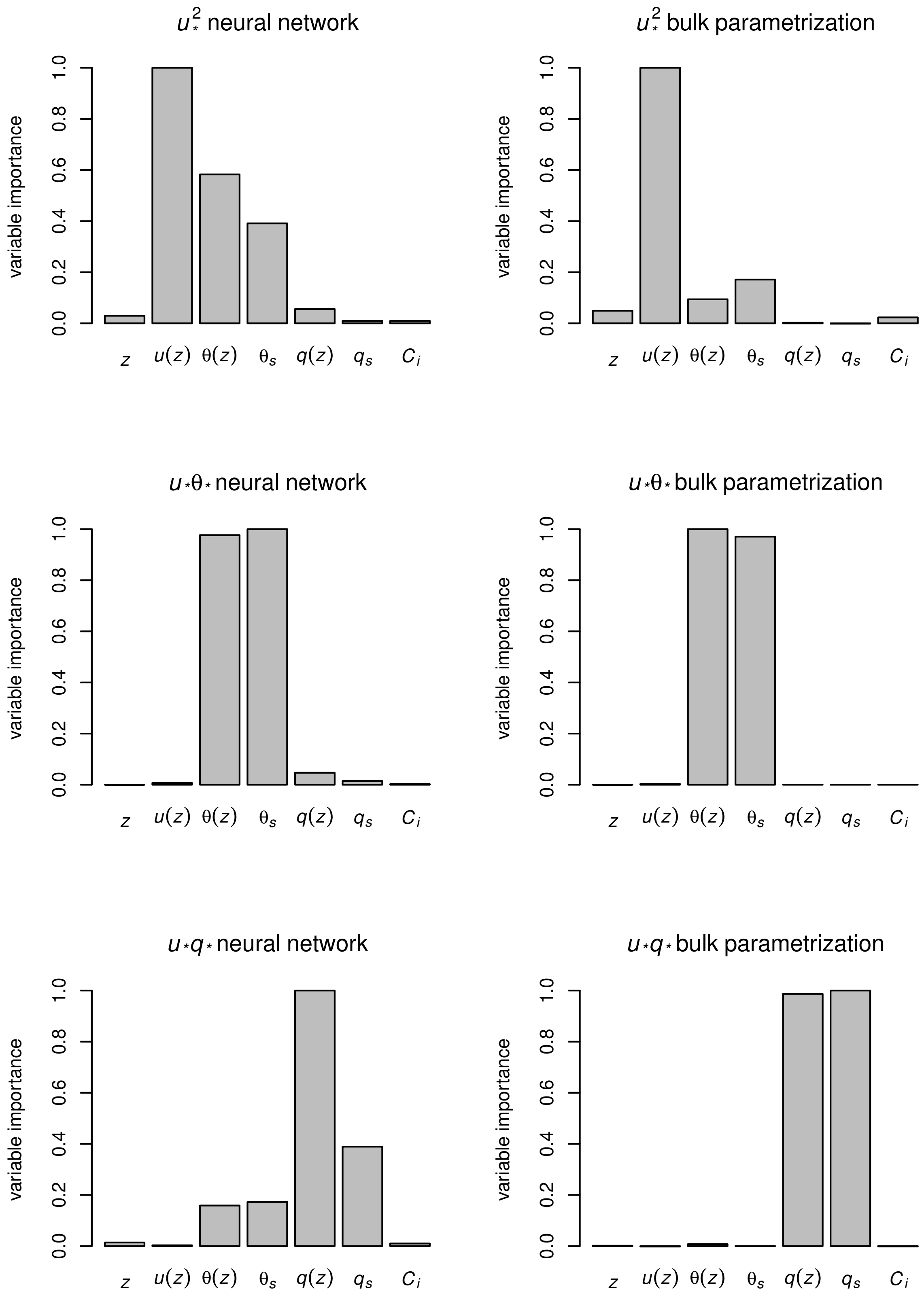

Figure 3: Bar plots showing the relative importance of the different input variables to the bulk algorithm and neural-network models. Variable importance metrics were computed using the "feature importance" algorithm of Breiman (2001).

The neural-network parametrization of the sensible heat flux $u_\star\theta_\star$ was generally successful, yielding reductions in RMSE of 12 % for the AO16 data and for two of the SHEBA PAM stations. For the ACSE dataset, and to a far greater extent ACCACIA, performance of the neural networks was worse. This is again a consequence of extrapolation failure. In ACCACIA, not only do all observations lie outside the range of z in the training set, but the observed fluxes themselves are often of much greater magnitude: the largest value of $u_\star\theta_\star$ in ACCACIA is more than five times larger than anything seen elsewhere. In both ACCACIA and ACSE, the neural networks systematically underestimate the magnitude of the extreme fluxes. This is because the use of a sigmoid activation function means the neural-network predictions will extrapolate roughly linearly until such a point as all the sigmoid activations on all the relevant neurons are saturated, after which the predictions will be insensitive to more extreme data. In this sense, the neural-network predictions in unseen conditions may be considered conservative estimates, as can be seen in Figure 2. On the other hand, the bulk algorithm suffered no such intrinsic limitation in the magnitude of its flux predictions. The high correlation between the observed and bulk-estimated fluxes in ACCACIA and ACSE is largely due to co-occurrence of large positive or negative fluxes. There was an anomalously low correlation between observed and modelled (both bulk and neural-network) fluxes at the SHEBA Cleveland-Seattle-Maui PAM station. The Seattle site in particular was surrounded by ice ridging and the PAM station was downwind of a pressure ridge that disturbed the turbulence (Andreas et al., 2010a), so conditions there were not representative of those at the other sites.

Unlike the parametrizations of $u_\star^2$ and $u_\star\theta_\star$, the neural-network estimator of the latent heat flux $u_\star q_\star$ did not deliver performance improvements over the bulk algorithm, although for ACSE and AO16 performance of the two methods was similar. It should be noted that observed fluxes of latent heat are not available at the SHEBA PAM stations, rendering the training sets in the cross-validation experiments both smaller and less representative of the corresponding validation sets. Latent heat flux observations in the Arctic are comparatively scarce, and of those available in this study the SHEBA tower contributes 94 %. The range of flux magnitudes observed at the SHEBA tower is small in comparison to the other datasets. Encouragingly, where the training sets are representative of conditions in the validation set (ACSE and AO16), performance of the neural networks is on par with that of the bulk. Given some of the performance gains of the neural-network parametrizations of $u_\star^2$ and $u_\star\theta_\star$, we might reasonably expect a training set which samples densely a wider range of flux values to enable an improved latent heat flux parametrization. Note, however, that the bulk algorithm generally performs worse for $u_\star q_\star$ than for the other fluxes, a finding consistent with results obtained outside the polar regions (e.g. McCandless et al., 2022). For example, it systematically under-predicts those fluxes with larger magnitudes.

Results from the variable importance analysis illustrate some of the differences in how the bulk algorithm and neural networks make use of their respective inputs (see Figure 3). For $u_\star^2$, the primary difference is that the neural-network model is more sensitive to the temperature inputs, as well as to the air humidity. The differences for $u_\star\theta_\star$ are less pronounced; however, the neural-network model is relatively more sensitive to the non-temperature inputs. For $u_\star q_\star$, the neural networks are less sensitive to surface humidity than to air humidity, but also have a stronger dependence on the temperature inputs. Due to the way it is constructed, the variable importance metric can be interpreted as a measure of robustness to error in the various inputs. It is therefore unsurprising that the accuracy of the bulk algorithm depends critically on the relevant gradient variables for each flux. Note that the variable importance metric is strictly relative, so even a very low score does not necessarily imply that an input is uninformative in an absolute sense. Perhaps because they are trained on noisy observational data, the neural networks appear to possess more redundancy across important inputs. For example, in the case of $u_\star q_\star$, surface temperature θ_s is an excellent proxy variable from which to estimate surface humidity q_s.

5 Summary and future directions

As the rapid melting of Arctic sea ice continues unabated, and as we become increasingly aware of the serious consequences of warming in the polar regions, the need for accurate representations of the relevant heat-transfer processes in climate models is greater than ever. Surface turbulent fluxes are a key mechanism for heat transfer between the atmosphere and ocean / sea ice, and yet their parametrization in current-generation climate models is based on traditional bulk formulae, originally calibrated in the tropics and mid-latitudes. Although polar-specific bulk formulations have been developed, their adoption in GCMs has been limited, in part due to the small number of field observations against which to validate their performance, and also because of difficulties modelling stable boundary layers generally.

The data-driven or machine-learning approach to parametrization of surface turbulent fluxes has emerged in recent years as an alternative or complement to bulk algorithms. In this study, it has been proposed to encode in a machine-learning model the relationships observed in practice between surface fluxes and meteorological predictor variables in the Arctic. To investigate feasibility of the data-driven approach, we have trained neural-network models using a database of observations assembled from several Arctic field campaigns. A bulk-algorithm implementation, containing advanced polar-specific parametrizations from the literature, has been used as a benchmark against which to test performance of machine-learning models.

Using a cross-validation scheme, the out-of-sample predictive accuracies of the bulk and neural-network parametrizations have been objectively compared. The neural-network parametrizations of the momentum and sensible heat fluxes were found to match and in some cases outperform their bulk counterparts in a RMSE sense. However, the neural-network latent heat flux parametrization was less successful and was generally outperformed by the bulk algorithm, probably due to insufficiently informative training data.

These results are encouraging and suggest directions for future research. Firstly, it may be possible to improve performance of the data-driven latent heat flux parametrization using observations from the recent MOSAiC campaign, where a large volume of data has been collected (Shupe et al., 2022). We would expect an expanded training dataset containing more extreme values to improve the representation of large fluxes. With the MOSAiC data, it may be possible to investigate sea ice surface characteristics (e.g. ridges, level ice, refrozen leads) as potential inputs to our algorithm. Another question to investigate is whether the bulk parametrization of latent heat flux can be used to inform or constrain neural-network training in a hybrid parametrization approach: incorporation of physical constraints has already been found to improve the extrapolation capability of machine-learning models (Zhao et al., 2019). There is also the question of whether results obtained here can be extended to sea ice in the Southern Ocean. Finally, having already obtained promising data-driven parametrizations of momentum and sensible heat flux, an immediate next step is to implement these models in a GCM. This step will likely be non-trivial: due to compensating errors, better agreement with in-situ observations is no guarantee that a new turbulence parametrization will improve predictive skill of the GCM (Sandu et al., 2013). Implementation in a GCM may benefit from *equation discovery* techniques (e.g. AI Feynman, Udrescu and Tegmark, 2020), whereby explicit equations are found which approximate well the neural networks. Explicit equations could reduce computational cost and would enable easier diagnostics in the case of model crashes. Sensitivity studies with a GCM will allow us to assess the impact of these new parametrizations on the polar atmosphere and on the melting of sea ice.

Acknowledgements

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.

The SHEBA data were provided by NCAR/EOL under the sponsorship of the National Science Foundation. We gratefully acknowledge the SHEBA Atmospheric Surface Flux Group (ASFG), who were responsible for the surface flux measurements during the

SHEBA project (E. L. Andreas, C. W. Fairall, P. S. Guest and P. O. G. Persson).

The neural-network architecture schematic in Figure 1 was created using the `plotnet` function in the *NeuralNetTools* package for R (Beck, 2018). Fitting of neural-network ensembles was automated with the help of wrapper functions in the R package *caret* (Kuhn and Johnson, 2013). Bootstrapped significance testing of performance-metric differences was performed using the R package *boot* (Canty and Ripley, 2022; Davison and Hinkley, 1997).

We thank the three anonymous reviewers for their thoughtful comments and suggestions, which helped improve the manuscript.

Declarations

Ethical approval

This declaration is not applicable.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

IMB, IAR, ADE and JP provided the ACCACIA, ACSE and AO16 data. VG built a local database from all the observational campaigns and implemented the bulk algorithm in Python. VG and SB advised on the bulk methodology and physical choices. DPC carried out all the analyses. DPC wrote the article with VG. All authors reviewed and provided feedback on the manuscript.

Funding

This work was supported by a national funding by the Agence Nationale de la Recherche within the framework of the Investissement d'Avenir programme under the ANR-17-

MPGA-0003 ASET 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).

The ACCACIA field campaign was supported by the UK Natural Environment Research Council (NERC) grant numbers NE/I028653/1, NE/I028858/1, and NE/I028297/1.

The participation of Ian Brooks and John Prytherch in the ACSE field campaign was supported by NERC grant number NE/K011820/1.

The AO16 measurements were supported by the the Swedish Polar Research Secretariat. John Prytherch was also supported by the Knut and Alice Wallenberg Foundation (grant number 2016-0024).

Availability of data and materials

Code availability

The polar-specific bulk algorithm used in this study and described in Section 3.1 is available to download as a Python library from GitHub (`https://github.com/virginieguemas/CDlib`).

Data availability

ACCACIA flight data are available from the CEDA archive:

- `https://doi.org/10.5285/0844186db1ba9e20319a2560f8d61651` (MASIN);
- `https://catalogue.ceda.ac.uk/uuid/c064b0c150274a1cbd18c563573f392e` (FAAM).

ACSE cruise data are available from the CEDA archive (`https://doi.org/10.5285/c6f1b1ff16f8407386e2d643bc5b916a`, Brooks et al., 2022a).

AO16 cruise data are available from the CEDA archive (`https://doi.org/10.5285/614752d35dc147a598d5421443fb50e8`, Brooks et al., 2022b).

SHEBA data are available from the NCAR Earth Observing Laboratory:

- (`https://doi.org/10.5065/D65H7DNS`, Andreas et al., 2007) (ASFG tower);
- (`https://doi.org/10.5065/D6ZC8170`, Andreas et al., 2012) (PAM stations).

NSIDC sea ice concentration data are available from the NSIDC archive (`https://doi.org/10.7265/efmz-2t65`, Meier et al., 2021).

A Polar-specific bulk parametrizations

A.1 Stability correction of Grachev et al. (2007)

Grachev et al. (2007) assumed the form of the vertical profiles to be:

$$a(z) = p_1(\ln z)^2 + p_2 \ln(z) + p_3, \tag{8}$$

where a can be the wind speed, potential temperature, specific humidity, or any other variable, and p_1, p_2, p_3 are constants to be determined. By taking the derivative of eqn (8), the vertical gradients can be obtained through a fit to observations. Using data from the SHEBA tower, Grachev et al. (2007) proposed the "SHEBA profile functions":

$$\phi_M = 1 + \frac{6.5\zeta(1+\zeta)^{\frac{1}{3}}}{1.3+\zeta}, \tag{9}$$

$$\phi_H = 1 + \frac{5\zeta + 5\zeta^2}{1 + 3\zeta + \zeta^2}, \tag{10}$$

where ζ is the Monin-Obukhov stability parameter and ϕ_M,ϕ_H are the non-dimensional stability profile functions for momentum and sensible/latent heat respectively.

A.2 Aerodynamic roughness model of Andreas et al. (2010b)

Based on the SHEBA winter data, Andreas et al. (2010b) concluded that the aerodynamic roughness z_0 did not significantly depend on the atmospheric stability. They proposed the following unified parametrization:

$$z_0 = 0.135\frac{\nu}{u_\star} + B\tanh^3(13u_\star), \tag{11}$$

where ν is the kinematic viscosity of air in m^{2}s^{-1} and $B = 2.3\times10^{-4}$. The first term on the right models aerodynamically smooth regimes, while the second treats aerodynamically rough regimes as well as the transition from smooth to rough flows.

A.3 Scalar roughness model of Andreas (1987)

Andreas (1987) proposed modelling the ratio of the scalar and aerodynamic roughness lengths as a function of the roughness Reynolds number:

$$\ln\frac{z_s}{z_0} = b_{0,s} + b_{1,s}\ln R_\star + b_{2,s}(\ln R_\star)^2, \tag{12}$$

where z_s is the scalar roughness for temperature ($s = T$) or humidity ($s = Q$) and $R_\star = \frac{u_\star z_0}{\nu}$ is the roughness Reynolds number. Polynomial coefficients $b_{i,s}$ are tabulated for smooth ($R_\star \leq 0.135$), transitional ($0.135 < R_\star < 2.5$) and rough ($2.5 \leq R_\star < 1000$) surfaces in Andreas (1987).

A.4 Form drag parametrization

The form drag parametrization used in this study assumes that the drag coefficient C_D in eqn (1) (the bulk exchange coefficient for momentum) can be partitioned as:

$$C_D = C_{D,w}(1 - C_i) + C_{D,i}C_i + C_{D,f}, \tag{13}$$

where $C_{D,w}$ and $C_{D,i}$ denote the skin drag coefficients over water and ice respectively, C_i is the sea ice concentration, and $C_{D,f}$ is the form drag contribution. We obtain $C_{D,f}$ by applying a stability correction to its neutral counterpart $C_{DN,f}$:

$$C_{DN,f} = C_{DN,f,w}(1 - C_i) + C_{DN,f,i}C_i, \tag{14}$$

where $C_{DN,f,k}$ are form-induced drag coefficients over water ($k = w$) and ice ($k = i$). Following Lüpkes et al. (2012) and Lüpkes and Gryanik (2015), we use:

$$C_{DN,f,k} = \frac{c_e}{2}\left[\frac{\ln\dfrac{h_{fc}}{z_{0,k}}}{\ln\dfrac{10}{z_{0,k}}}\right]^2 \frac{h_{fc}}{D}(1 - C_i)^{\beta}C_i, \tag{15}$$

where $c_e = 0.4$ is the effective resistance coefficient, $h_{fc} = 0.41$ m is the ice floe freeboard, $z_{0,k}$ is the skin drag aerodynamic roughness over water/ice, $D = 8$ m is the average diameter of leads or melt ponds and $\beta = 1$ is an optional exponent.

References

Andreas, E. L. (1987). A theory for the scalar roughness and the scalar transfer coefficients over snow and sea ice. *Boundary-Layer Meteorology*, 38(1):159–184.

Andreas, E. L. (1998). The Atmospheric Boundary Layer Over Polar Marine Surfaces. In Leppäranta, M., editor, *Physics of Ice-Covered Seas*, volume 2, pages 715–773. Helsinki University Press, Helsinki.

Andreas, E. L. (2011). A relationship between the aerodynamic and physical roughness of winter sea ice. *Quarterly Journal of the Royal Meteorological Society*, 137(659):1581–1588.

Andreas, E. L., Claffey, K. J., Jordan, R. E., Fairall, C. W., Guest, P. S., Persson, P. O. G., and Grachev, A. A. (2006). Evaluations of the von Kármán constant in the atmospheric surface layer. *Journal of Fluid Mechanics*, 559:117–149.

Andreas, E. L., Fairall, C., Guest, P., and Persson, O. (2007). Tower, 5-level hourly measurements plus radiometer and surface data at Met City (ASFG). Version 1.0.

Andreas, E. L., Fairall, C., Guest, P., and Persson, O. (2012). Ice Camp Surface Mesonet NCAR PAM-III 1 hour (FINAL). Version 1.0.

Andreas, E. L., Fairall, C. W., Grachev, A. A., Guest, P. S., Horst, T. W., Jordan, R. E., and Persson, P. O. G. (2003). Turbulent transfer coefficients and roughness lengths over sea ice: The SHEBA results. In *7th Conference on Polar Meteorology and Oceanography. Hyannis, MA, Amer. Meteorol. Soc., Proc.*

Andreas, E. L., Fairall, C. W., Guest, P. S., and Persson, P. O. G. (1999). An overview of the SHEBA atmospheric surface flux program. In *13th Symposium on Boundary Layers and Turbulence. Dallas, TX, Amer. Meteorol. Soc., Proc.*, pages 550–555.

Andreas, E. L., Guest, P. S., Persson, P. O. G., Fairall, C. W., Horst, T. W., Moritz, R. E., and Semmer, S. R. (2002). Near-surface water vapor over polar sea ice is always near ice saturation. *Journal of Geophysical Research: Oceans*, 107(C10):SHE 8–1–SHE 8–15.

Andreas, E. L., Horst, T. W., Grachev, A. A., Persson, P. O. G., Fairall, C. W., Guest, P. S., and Jordan, R. E. (2010a). Parametrizing turbulent exchange over summer sea ice and the marginal ice zone. *Quarterly Journal of the Royal Meteorological Society*, 136(649):927–943.

Andreas, E. L., Persson, P. O. G., Grachev, A. A., Jordan, R. E., Horst, T. W., Guest, P. S., and Fairall, C. W. (2010b). Parameterizing Turbulent Exchange over Sea Ice in Winter. *Journal of Hydrometeorology*, 11(1):87–104.

Audouin, O., Roehrig, R., Couvreux, F., and Williamson, D. (2021). Modeling the GABLS4 Strongly-Stable Boundary Layer With a GCM Turbulence Parameterization: Parametric Sensitivity or Intrinsic Limits? *Journal of Advances in Modeling Earth Systems*, 13(3):e2020MS002269.

Baas, P., Steeneveld, G. J., van de Wiel, B. J. H., and Holtslag, A. a. M. (2006). Exploring Self-Correlation in Flux–Gradient Relationships for Stably Stratified Conditions. *Journal of the Atmospheric Sciences*, 63(11):3045–3054.

Beck, M. W. (2018). NeuralNetTools: Visualization and Analysis Tools for Neural Networks. *Journal of Statistical Software*, 85:1–20.

Beljaars, A. C. M. and Holtslag, A. a. M. (1991). Flux Parameterization over Land Surfaces for Atmospheric Models. *Journal of Applied Meteorology and Climatology*, 30(3):327–341.

Bonan, D. B., Lehner, F., and Holland, M. M. (2021a). Partitioning uncertainty in projections of Arctic sea ice. *Environmental Research Letters*, 16(4):044002.

Bonan, D. B., Schneider, T., Eisenman, I., and Wills, R. C. J. (2021b). Constraining the Date of a Seasonally Ice-Free Arctic Using a Simple Model. *Geophysical Research Letters*, 48(18):e2021GL094309.

Bosveld, F. C., Baas, P., Steeneveld, G.-J., Holtslag, A. A. M., Angevine, W. M., Bazile, E., de Bruijn, E. I. F., Deacu, D., Edwards, J. M., Ek, M., Larson, V. E., Pleim, J. E., Raschendorfer, M., and Svensson, G. (2014). The Third GABLS Intercomparison Case for Evaluation Studies of Boundary-Layer Models. Part B: Results and Process Understanding. *Boundary-Layer Meteorology*, 152(2):157–187.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Brooks, I. M., Prytherch, J., and Srivastava, P. (2022a). CANDIFLOS : Surface fluxes from ACSE measurement campaign on icebreaker Oden, 2014.

Brooks, I. M., Prytherch, J., and Srivastava, P. (2022b). CANDIFLOS : Surface fluxes from AO2016 measurement campaign on icebreaker Oden, 2014.

Broyden, C. G. (1970). The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90.

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

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery.

Cohen, J., Screen, J. A., Furtado, J. C., Barlow, M., Whittleston, D., Coumou, D., Francis, J., Dethloff, K., Entekhabi, D., Overland, J., and Jones, J. (2014). Recent Arctic amplification and extreme mid-latitude weather. *Nature Geoscience*, 7(9):627–637.

Cohen, J., Zhang, X., Francis, J., Jung, T., Kwok, R., Overland, J., Ballinger, T. J., Bhatt, U. S., Chen, H. W., Coumou, D., Feldstein, S., Gu, H., Handorf, D., Henderson, G., Ionita, M., Kretschmer, M., Laliberte, F., Lee, S., Linderholm, H. W., Maslowski, W., Peings, Y., Pfeiffer, K., Rigor, I., Semmler, T., Stroeve, J., Taylor, P. C., Vavrus, S., Vihma, T., Wang, S., Wendisch, M., Wu, Y., and Yoon, J. (2020). Divergent consensuses on Arctic amplification influence on midlatitude severe winter weather. *Nature Climate Change*, 10(1):20–29.

Cuxart, J., Holtslag, A. A. M., Beare, R. J., Bazile, E., Beljaars, A., Cheng, A., Conangla, L., Ek, M., Freedman, F., Hamdi, R., Kerstein, A., Kitagawa, H., Lenderink, G., Lewellen, D., Mailhot, J., Mauritsen, T., Perov, V., Schayes, G., Steeneveld, G.-J., Svensson, G., Taylor, P., Weng, W., Wunsch, S., and Xu, K.-M. (2006). Single-Column Model Intercomparison for a Stably Stratified Atmospheric Boundary Layer. *Boundary-Layer Meteorology*, 118(2):273–303.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Edson, J. B., Jampana, V., Weller, R. A., Bigorre, S. P., Plueddemann, A. J., Fairall,

C. W., Miller, S. D., Mahrt, L., Vickers, D., and Hersbach, H. (2013). On the Exchange of Momentum over the Open Ocean. *Journal of Physical Oceanography*, 43(8):1589–1610.

Edson, J. B., Zappa, C. J., Ware, J. A., McGillis, W. R., and Hare, J. E. (2004). Scalar flux profile relationships over the open ocean. *Journal of Geophysical Research: Oceans*, 109(C8).

Elvidge, A. D., Renfrew, I., Edwards, J., Brooks, I., Srivastava, P., and Weiss, A. (2023). Improved simulation of the polar atmospheric boundary layer by accounting for aerodynamic roughness in the parameterisation of surface scalar exchange over sea ice. *Journal of Advances in Modeling Earth Systems*.

Elvidge, A. D., Renfrew, I. A., Brooks, I. M., Srivastava, P., Yelland, M. J., and Prytherch, J. (2021). Surface Heat and Moisture Exchange in the Marginal Ice Zone: Observations and a New Parameterization Scheme for Weather and Climate Models. *Journal of Geophysical Research: Atmospheres*, 126(17):e2021JD034827.

Elvidge, A. D., Renfrew, I. A., Weiss, A. I., Brooks, I. M., Lachlan-Cope, T. A., and King, J. C. (2016). Observations of surface momentum exchange over the marginal ice zone and recommendations for its parametrisation. *Atmospheric Chemistry and Physics*, 16(3):1545–1563.

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

Fiedler, E. K., Lachlan-Cope, T. A., Renfrew, I. A., and King, J. C. (2010). Convective heat transfer over thin ice covered coastal polynyas. *Journal of Geophysical Research: Oceans*, 115(C10).

Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Garratt, J. R. (1994). *The Atmospheric Boundary Layer.* Cambridge Atmospheric and Space Science Series. Cambridge University Press, Cambridge, UK.

Gascard, J.-C., Zhang, J., and Rafizadeh, M. (2019). Rapid decline of Arctic sea ice volume: Causes and consequences. *The Cryosphere Discussions*, pages 1–29.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26.

Grachev, A. A., Andreas, E. L., Fairall, C. W., Guest, P. S., and Persson, P. O. G. (2007). SHEBA flux–profile relationships in the stable atmospheric boundary layer. *Boundary-Layer Meteorology*, 124(3):315–333.

Grachev, A. A. and Fairall, C. W. (1997). Dependence of the Monin–Obukhov Stability Parameter on the Bulk Richardson Number over the Ocean. *Journal of Applied Meteorology and Climatology*, 36(4):406–414.

Grachev, A. A., Fairall, C. W., and Bradley, E. F. (2000). Convective Profile Constants Revisited. *Boundary-Layer Meteorology*, 94(3):495–515.

Grachev, A. A., Fairall, C. W., Persson, P. O. G., Andreas, E. L., and Guest, P. S. (2002). Stable boundary-layer regimes observed during the SHEBA Experiment. In *15th Symposium on Boundary Layers and Turbulence. Wageningen, The Netherlands, Amer. Meteorol. Soc., Proc.*, pages 374–377.

Grachev, A. A., Fairall, C. W., Persson, P. O. G., Andreas, E. L., and Guest, P. S. (2005). Stable Boundary-Layer Scaling Regimes: The SHEBA Data. *Boundary-Layer Meteorology*, 116(2):201–235.

Graversen, R. G., Mauritsen, T., Tjernström, M., Källén, E., and Svensson, G. (2008). Vertical structure of recent Arctic warming. *Nature*, 451(7174):53–56.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, 1(3):297–310.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). Neural Networks. In Hastie, T., Tibshirani, R., and Friedman, J., editors, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, pages 389–416. Springer, New York, NY.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Jung, T., Doblas-Reyes, F., Goessling, H., Guemas, V., Bitz, C., Buontempo, C., Caballero, R., Jakobson, E., Jungclaus, J., Karcher, M., Koenigk, T., Matei, D., Overland, J., Spengler, T., and Yang, S. (2015). Polar Lower-Latitude Linkages and Their Role in Weather and Climate Prediction. *Bulletin of the American Meteorological Society*, 96(11):ES197–ES200.

Jung, T., Gordon, N. D., Bauer, P., Bromwich, D. H., Chevallier, M., Day, J. J., Dawson, J., Doblas-Reyes, F., Fairall, C., Goessling, H. F., Holland, M., Inoue, J., Iversen, T., Klebe, S., Lemke, P., Losch, M., Makshtas, A., Mills, B., Nurmi, P., Perovich, D., Reid, P., Renfrew, I. A., Smith, G., Svensson, G., Tolstykh, M., and Yang, Q. (2016). Advancing Polar Prediction Capabilities on Daily to Seasonal Time Scales. *Bulletin of the American Meteorological Society*, 97(9):1631–1647.

King, J. C., Lachlan-Cope, T. A., Ladkin, R. S., and Weiss, A. (2008). Airborne Measurements in the Stable Boundary Layer over the Larsen Ice Shelf, Antarctica. *Boundary-Layer Meteorology*, 127(3):413–428.

Kovacs, K. M., Lydersen, C., Overland, J. E., and Moore, S. E. (2011). Impacts of changing sea-ice conditions on Arctic marine mammals. *Marine Biodiversity*, 41(1):181–194.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer, New York, NY.

Lesins, G., Duck, T. J., and Drummond, J. R. (2012). Surface Energy Balance Framework for Arctic Amplification of Climate Change. *Journal of Climate*, 25(23):8277–8288.

Leufen, L. H. and Schädler, G. (2019). Calculating the turbulent fluxes in the atmospheric surface layer with neural networks. *Geoscientific Model Development*, 12(5):2033–2047.

Liu, J., Song, M., Zhu, Z., Horton, R. M., Hu, Y., and Xie, S.-P. (2022). Arctic sea-ice loss is projected to lead to more frequent strong El Niño events. *Nature Communications*, 13(1):4952.

Lüpkes, C. and Gryanik, V. M. (2015). A stability-dependent parametrization of transfer coefficients for momentum and heat over polar sea ice to be used in climate models. *Journal of Geophysical Research: Atmospheres*, 120(2):552–581.

Lüpkes, C., Gryanik, V. M., Hartmann, J., and Andreas, E. L. (2012). A parametrization, based on sea ice morphology, of the neutral atmospheric drag coefficients for weather prediction and climate models. *Journal of Geophysical Research: Atmospheres*, 117(D13).

Lüpkes, C., Gryanik, V. M., Rösel, A., Birnbaum, G., and Kaleschke, L. (2013). Effect of sea ice morphology during Arctic summer on atmospheric drag coefficients used in climate models. *Geophysical Research Letters*, 40(2):446–451.

McCandless, T., Gagne, D. J., Kosović, B., Haupt, S. E., Yang, B., Becker, C., and Schreck, J. (2022). Machine Learning for Improving Surface-Layer-Flux Estimates. *Boundary-Layer Meteorology*, 185(2):199–228.

Meier, W. N., Fetterer, F., Windnagel, A., and Stewart, S. (2021). NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration, Version 4.

Meier, W. N., Hovelsrud, G. K., van Oort, B. E., Key, J. R., Kovacs, K. M., Michel, C., Haas, C., Granskog, M. A., Gerland, S., Perovich, D. K., Makshtas, A., and Reist, J. D. (2014). Arctic sea ice in transformation: A review of recent observed changes and impacts on biology and human activity. *Reviews of Geophysics*, 52(3):185–217.

Monin, A. S. and Obukhov, A. M. (1954). Basic laws of turbulent mixing in the surface layer of the atmosphere. *Tr. Akad. Nauk SSSR Geophiz. Inst.*, 24(151):163–187.

Muñoz-Esparza, D., Becker, C., Sauer, J. A., Gagne II, D. J., Schreck, J., and Kosović, B. (2022). On the Application of an Observations-Based Machine Learning Parameterization of Surface Layer Fluxes Within an Atmospheric Large-Eddy Simulation Model. *Journal of Geophysical Research: Atmospheres*, 127(16):e2021JD036214.

Notz, D., Haumann, F. A., Haak, H., Jungclaus, J. H., and Marotzke, J. (2013). Arctic sea-ice evolution as modeled by Max Planck Institute for Meteorology's Earth system model. *Journal of Advances in Modeling Earth Systems*, 5(2):173–194.

Pelliccioni, A., Poli, U., Agnello, P., and Coni, A. (1999). Application of neural networks to model the Monin-Obukhov length and the mixed-layer height from ground-based meteorological data. *Transactions on Ecology and the Environment*, 29:1055–1064.

Persson, O., Shupe, M. D., Tjernström, M., Sedlar, J., Brooks, I. M., Brooks, B. J., Bjork, G., Prytherch, J., Salisbury, D., Achtert, P., Sotiropoulou, G., Johnston, P. E., and Wolfe, D. (2015). ATMOSPHERE-ICE-OCEAN INTERACTIONS DURING SUMMER MELT AND EARLY AUTUMN FREEZE-UP: OBSERVATIONS FROM THE ACSE FIELD PROGRAM.

Persson, P. O. G., Fairall, C. W., Andreas, E. L., Guest, P. S., and Perovich, D. K. (2002). Measurements near the Atmospheric Surface Flux Group tower at SHEBA: Near-surface conditions and surface energy budget. *Journal of Geophysical Research: Oceans*, 107(C10):SHE 21–1–SHE 21–35.

Petersen, G. N. and Renfrew, I. A. (2009). Aircraft-based observations of air–sea fluxes over Denmark Strait and the Irminger Sea during high wind speed conditions. *Quarterly Journal of the Royal Meteorological Society*, 135(645):2030–2045.

Post, E., Bhatt, U. S., Bitz, C. M., Brodie, J. F., Fulton, T. L., Hebblewhite, M., Kerby, J., Kutz, S. J., Stirling, I., and Walker, D. A. (2013). Ecological Consequences of Sea-Ice Decline. *Science*, 341(6145):519–524.

Previdi, M., Smith, K. L., and Polvani, L. M. (2021). Arctic amplification of climate change: A review of underlying mechanisms. *Environmental Research Letters*, 16(9):093003.

Prytherch, J., Brooks, I. M., Crill, P. M., Thornton, B. F., Salisbury, D. J., Tjernström, M., Anderson, L. G., Geibel, M. C., and Humborg, C. (2017). Direct determination of the air-sea CO2 gas transfer velocity in Arctic sea ice regions. *Geophysical Research Letters*, 44(8):3770–3778.

Qin, Z., Su, G.-l., Yu, Q., Hu, B.-m., and Li, J. (2005a). Modeling water and carbon fluxes above summer maize field in North China Plain with back-propagation neural networks. *Journal of Zhejiang University-SCIENCE B*, 6(5):418–426.

Qin, Z., Yu, Q., Li, J., Wu, Z.-y., and Hu, B.-m. (2005b). Application of least squares vector machines in modelling water vapor and carbon dioxide fluxes over a cropland. *Journal of Zhejiang University-SCIENCE B*, 6(6):491–495.

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Renfrew, I. A., Elvidge, A. D., and Edwards, J. M. (2019). Atmospheric sensitivity to marginal-ice-zone drag: Local and global responses. *Quarterly Journal of the Royal Meteorological Society*, 145(720):1165–1179.

Renfrew, I. A., Moore, G. W. K., Kristjánsson, J. E., Ólafsson, H., Gray, S. L., Petersen, G. N., Bovis, K., Brown, P. R. A., Føre, I., Haine, T., Hay, C., Irvine, E. A., Lawrence, A., Ohigashi, T., Outten, S., Pickart, R. S., Shapiro, M., Sproson, D., Swinbank, R., Woolley, A., and Zhang, S. (2008). THE GREENLAND FLOW DISTORTION EXPERIMENT. *Bulletin of the American Meteorological Society*, 89(9):1307–1324.

Ripley, B. D. (1996). Feed-forward Neural Networks. In *Pattern Recognition and Neural Networks*, pages 143–180. Cambridge University Press, Cambridge.

Rothrock, D. A., Yu, Y., and Maykut, G. A. (1999). Thinning of the Arctic sea-ice cover. *Geophysical Research Letters*, 26(23):3469–3472.

Safa, B., Arkebauer, T. J., Zhu, Q., Suyker, A., and Irmak, S. (2018). Latent heat and sensible heat flux simulation in maize using artificial neural networks. *Computers and Electronics in Agriculture*, 154:155–164.

Sandu, I., Beljaars, A., Bechtold, P., Mauritsen, T., and Balsamo, G. (2013). Why is it so difficult to represent stably stratified conditions in numerical weather prediction (NWP) models? *Journal of Advances in Modeling Earth Systems*, 5(2):117–133.

Screen, J. A. and Simmonds, I. (2010). The central role of diminishing sea ice in recent Arctic temperature amplification. *Nature*, 464(7293):1334–1337.

Serreze, M. C., Barrett, A. P., Stroeve, J. C., Kindig, D. N., and Holland, M. M. (2009). The emergence of surface-based Arctic amplification. *The Cryosphere*, 3(1):11–19.

Serreze, M. C. and Barry, R. G. (2011). Processes and impacts of Arctic amplification: A research synthesis. *Global and Planetary Change*, 77(1):85–96.

Serreze, M. C. and Francis, J. A. (2006). The Arctic Amplification Debate. *Climatic Change*, 76(3):241–264.

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656.

Shupe, M. D., Rex, M., Blomquist, B., Persson, P. O. G., Schmale, J., Uttal, T., Althausen, D., Angot, H., Archer, S., Bariteau, L., Beck, I., Bilberry, J., Bucci, S., Buck, C., Boyer, M., Brasseur, Z., Brooks, I. M., Calmer, R., Cassano, J., Castro, V., Chu, D., Costa, D., Cox, C. J., Creamean, J., Crewell, S., Dahlke, S., Damm, E., de Boer, G., Deckelmann, H., Dethloff, K., Dütsch, M., Ebell, K., Ehrlich, A., Ellis, J., Engelmann, R., Fong, A. A., Frey, M. M., Gallagher, M. R., Ganzeveld, L., Gradinger, R., Graeser, J., Greenamyer, V., Griesche, H., Griffiths, S., Hamilton, J., Heinemann, G., Helmig, D., Herber, A., Heuzé, C., Hofer, J., Houchens, T., Howard, D., Inoue, J., Jacobi, H.-W., Jaiser, R., Jokinen, T., Jourdan, O., Jozef, G., King, W., Kirchgaessner, A., Klingebiel, M., Krassovski, M., Krumpen, T., Lampert, A., Landing, W., Laurila, T., Lawrence, D., Lonardi, M., Loose, B., Lüpkes, C., Maahn, M., Macke, A., Maslowski, W., Marsay, C., Maturilli, M., Mech, M., Morris, S., Moser, M., Nicolaus, M., Ortega, P., Osborn, J., Pätzold, F., Perovich, D. K., Petäjä, T., Pilz, C., Pirazzini, R., Posman, K., Powers, H., Pratt, K. A., Preußer, A., Quéléver, L., Radenz, M., Rabe, B., Rinke, A., Sachs, T., Schulz, A., Siebert, H., Silva, T., Solomon, A., Sommerfeld, A., Spreen, G., Stephens, M., Stohl, A., Svensson, G., Uin, J., Viegas, J., Voigt, C., von der Gathen, P., Wehner, B., Welker, J. M., Wendisch, M., Werner, M., Xie, Z., and Yue, F. (2022). Overview of the MOSAiC expedition: Atmosphere. *Elementa: Science of the Anthropocene*, 10(1):00060.

Sotiropoulou, G., Tjernström, M., Sedlar, J., Achtert, P., Brooks, B. J., Brooks, I. M., Persson, P. O. G., Prytherch, J., Salisbury, D. J., Shupe, M. D., Johnston, P. E., and Wolfe, D. (2016). Atmospheric Conditions during the Arctic Clouds in Summer Experiment (ACSE): Contrasting Open Water and Sea Ice Surfaces during Melt and Freeze-Up Seasons. *Journal of Climate*, 29(24):8721–8744.

Srivastava, P., Brooks, I. M., Prytherch, J., Salisbury, D. J., Elvidge, A. D., Renfrew, I. A., and Yelland, M. J. (2022). Ship-based estimates of momentum transfer coefficient over sea ice and recommendations for its parameterization. *Atmospheric Chemistry and Physics*, 22(7):4763–4778.

Svensson, G., Holtslag, A. A. M., Kumar, V., Mauritsen, T., Steeneveld, G. J., Angevine, W. M., Bazile, E., Beljaars, A., de Bruijn, E. I. F., Cheng, A., Conangla, L., Cuxart, J., Ek, M., Falk, M. J., Freedman, F., Kitagawa, H., Larson, V. E., Lock, A., Mailhot, J., Masson, V., Park, S., Pleim, J., Söderberg, S., Weng, W., and Zampieri, M. (2011). Evaluation of the Diurnal Cycle in the Atmospheric Boundary Layer Over Land as Represented by a Variety of Single-Column Models: The Second GABLS Experiment. *Boundary-Layer Meteorology*, 140(2):177–206.

Thornton, B. F., Prytherch, J., Andersson, K., Brooks, I. M., Salisbury, D., Tjernström, M., and Crill, P. M. (2020). Shipborne eddy covariance observations of methane fluxes constrain Arctic sea emissions. *Science Advances*, 6(5):eaay7934.

Tjernström, M. and Jakobsson, M. (2021). Data from expedition Arctic Ocean, 2016.

Tynan, E. (2015). Effects of sea-ice loss. *Nature Climate Change*, 5(7):621–621.

Udrescu, S.-M. and Tegmark, M. (2020). AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631.

Uttal, T., Curry, J. A., McPhee, M. G., Perovich, D. K., Moritz, R. E., Maslanik, J. A., Guest, P. S., Stern, H. L., Moore, J. A., Turenne, R., Heiberg, A., Serreze, M. C., Wylie, D. P., Persson, O. G., Paulson, C. A., Halle, C., Morison, J. H., Wheeler, P. A., Makshtas, A., Welch, H., Shupe, M. D., Intrieri, J. M., Stamnes, K., Lindsey, R. W., Pinkel, R., Pegau, W. S., Stanton, T. P., and Grenfeld, T. C. (2002). Surface Heat Budget of the Arctic Ocean. *Bulletin of the American Meteorological Society*, 83(2):255–276.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S.* Statistics and Computing. Springer, New York, NY.

Vihma, T. (1995). Subgrid parameterization of surface heat and momentum fluxes over polar oceans. *Journal of Geophysical Research: Oceans*, 100(C11):22625–22646.

Vihma, T., Pirazzini, R., Fer, I., Renfrew, I. A., Sedlar, J., Tjernström, M., Lüpkes, C., Nygård, T., Notz, D., Weiss, J., Marsan, D., Cheng, B., Birnbaum, G., Gerland, S., Chechin, D., and Gascard, J. C. (2014). Advances in understanding and parameterization of small-scale physical processes in the marine Arctic climate system: A review. *Atmospheric Chemistry and Physics*, 14(17):9403–9450.

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., Colin, J., Guérémy, J.-F., Michou, M., Moine, M.-P., Nabat, P., Roehrig, R., Salas y Mélia, D., Séférian, R., Valcke, S., Beau, I., Belamari, S., Berthet, S., Cassou, C., Cattiaux, J., Deshayes, J., Douville, H., Ethé, C., Franchistéguy, L., Geoffroy, O., Lévy, C., Madec, G., Meurdesoif, Y., Msadek, R., Ribes, A., Sanchez-Gomez, E., Terray, L., and Waldman, R. (2019). Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, 11(7):2177–2213.

Wang, L., Zhang, Y., Yao, Y., Xiao, Z., Shang, K., Guo, X., Yang, J., Xue, S., and Wang, J. (2021). GBRT-Based Estimation of Terrestrial Latent Heat Flux in the Haihe River Basin from Satellite and Reanalysis Datasets. *Remote Sensing*, 13(6):1054.

Wang, X., Yao, Y., Zhao, S., Jia, K., Zhang, X., Zhang, Y., Zhang, L., Xu, J., and Chen, X. (2017). MODIS-Based Estimation of Terrestrial Latent Heat Flux over North America Using Three Machine Learning Algorithms. *Remote Sensing*, 9(12):1326.

Wulfmeyer, V., Pineda, J. M. V., Otte, S., Karlbauer, M., Butz, M. V., Lee, T. R., and Rajtschan, V. (2022). Estimation of the Surface Fluxes for Heat and Momentum in Unstable Conditions with Machine Learning and Similarity Approaches for the LAFE Data Set. *Boundary-Layer Meteorology*.

Xu, T., Guo, Z., Liu, S., He, X., Meng, Y., Xu, Z., Xia, Y., Xiao, J., Zhang, Y., Ma, Y., and Song, L. (2018). Evaluating Different Machine Learning Methods for Upscaling Evapotranspiration from Flux Towers to the Regional Scale. *Journal of Geophysical Research: Atmospheres*, 123(16):8674–8690.

Zhao, W. L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., Lin, C., Li, X., and Qiu, G. Y. (2019). Physics-Constrained Machine Learning of Evapotranspiration. *Geophysical Research Letters*, 46(24):14496–14507.