Geoscientific
Model Development

Model description paper

At-scale Model Output Statistics in mountain environments (AtsMOS v1.0)

Maximillian Van Wyk de Vries[1,2,3], **Tom Matthews**[4], **L. Baker Perry**[5], **Nirakar Thapa**[6], **and Rob Wilby**[7]

[1]Department of Geography, University of Cambridge, Cambridge CB2 3EL, UK
[2]Department of Earth Sciences, University of Cambridge, Cambridge CB3 0EZ, UK
[3]School of Geography and the Environment, University of Oxford, Oxford OX1 3QY, UK
[4]Department of Geography, Kings College London, London WC2B 4BG, UK
[5]Department of Geography and Planning, Appalachian State University, Boone, North Carolina, USA
[6]Department of Hydrology and Meteorology, Kathmandu, Nepal
[7]Department of Geography and Environment, Loughborough University, Loughborough, UK

Correspondence: Maximillian Van Wyk de Vries (msv27@cam.ac.uk)

Abstract. This paper introduces the AtsMOS (At-scale Model Output Statistics) workflow, designed to enhance mountain meteorology predictions through the downscaling of coarse numerical weather predictions using local observational data. AtsMOS provides a modular, open-source toolkit for local and large-scale forecasting of various meteorological variables through modified model output statistics – and may be applied to data from a single station or an entire network. We demonstrate its effectiveness through an example application at the summit of Mt. Everest, where it improves the prediction of both meteorological variables (e.g. wind speed, temperature) and derivative variables (e.g. facial frostbite time) critical for mountaineering safety. As a bridge between numerical weather prediction models and ground observations, AtsMOS contributes to hazard mitigation, water resource management, and other weather-dependent issues in mountainous regions and beyond.

1 Introduction

Accurate mountain weather forecasts facilitate improved hazard mitigation for the 300 million mountain inhabitants worldwide and contribute to effective sustainable resource management (e.g. Miner et al., 2020; Corbari et al., 2022). Furthermore, they are relevant to the 1.6 billion who live downstream of mountains and depend on their supply of freshwater or are susceptible to their hazards (Immerzeel et al., 2020). However, producing skilful forecasts in such environments is challenging. Major topographic variations cause large spatial variability in the weather, meaning that reality can diverge substantially from numerical weather prediction (NWP) grid-point forecasts within typical 10^2–10^3 km^2 grid cell areas (Zhang et al., 2022). Whilst consistent biases can be adjusted for (e.g. mismatches in elevation between forecast grid points and land surface locations of interest with knowledge of the lapse rate; Minder et al., 2010), the impact of unresolved processes – for instance, local valley or glacier winds driven by surface heat fluxes (Khadka et al., 2022) – is challenging to correct for a priori.

Although advances in NWP, such as finer grid resolutions and refinement of physical parameterization schemes, may enhance forecast performance in mountainous terrain, progress can be costly and slow (Bauer et al., 2015). A cheaper, faster, and more flexible option to improve forecasts for target locations is to statistically post-process NWP output through calibration to observations. Model output statistics (MOS), which applies multiple linear regression to adjust forecast fields, has historically been the most popular method in this regard (Glahn and Lowry, 1972; Glahn, 2014). In particular, MOS can be used to create forecasts of variables (predictands) not available in NWP model output (Rasp et al., 2020). Recent advancements in computational power have enabled machine learning to improve the performance

Table 1. Examples of weather variables (predictands) and sectors in which highly accurate, site-specific forecasts may be desirable.

Predictand	Example sector(s)
Precipitation amount and phase	Hazard forecasting (flood, avalanche); resource planning
Maximum wind gust	Aviation; mountaineering*; hazard (avalanche) forecasting
Ground temperature	(Road) transport; mountaineering
Wind chill temperature	Mountaineering
Cloud base and cloud top	Aviation; mountaineering
Probability of rime ice accretion	Communications
Facial frostbite time	Mountaineering

* The term mountaineering is used to represent a wider set of similar activities – e.g. hiking, skiing, and climbing.

of weather forecasts (Lam et al., 2023), including through post-processing (e.g. Lagerquist et al., 2017; Herman and Schumacher, 2018; Han et al., 2021; Grönquist et al., 2021).

However, practical barriers may limit the uptake of such developments at scale. For example, without reference workflows to facilitate the non-trivial task of accessing and pre-processing large NWP datasets, forecasts improved by machine learning are unlikely to reach the diverse range of potential end users in mountainous environments (Table 1). The benefits of highly accurate local weather predictions for use in other (e.g. hydrological) modelling chains may not be achieved if such forecasts are not made available in an interoperable format that follows well-known conventions, such as the "CF" – climate and forecast – convention (Eaton et al., 2023).

Hence, our paper aims to introduce a user-friendly, lightweight version of MOS to fill this gap. We describe modular Python code that calibrates and applies MOS, including state-of-the-art machine learning algorithms, to produce corrected forecasts in an interoperable format that can feed into other automated workflows to enable at-scale MOS. We anticipate that these features of AtsMOS (At-scale Model Output Statistics) will, combined with efforts to improve the availability of high-altitude weather observations worldwide (GEO Mountains 2022), offer a step change in the ability to forecast critical mountain weather variables.

In Sect. 2 we describe the main features of AtsMOS, before illustrating its use in forecasting the weather on the summit of Mt. Everest, where highly accurate predictions can be the difference between life and death (Sect. 3). In Sect. 4 we discuss opportunities and challenges in using AtsMOS more broadly.

2 The AtsMOS workflow

2.1 Workflow overview

AtsMOS is designed to be a computationally light and flexible template. It has (i) a flexible loading and pre-processing module, which draws in external data, deals with erroneous or missing data, and prepares it for further analysis. Our code
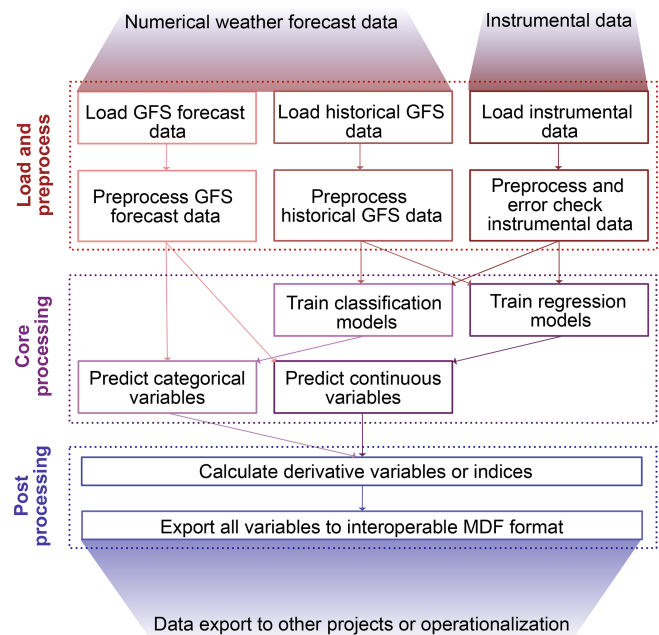
Figure 1. Overview diagram of the AtsMOS workflow. The loading of historical GFS data and loading and pre-processing of instrumental data are flexible components subject to user modification, while the others are fixed in this workflow.

here is intended as a template such that users may set up their own data loading and pre-processing as the need arises. We therefore do not describe each operation in detail but instead refer readers to the documented AtsMOS Jupyter notebook associated with this paper. It also has (ii) a core processing module, comprising a modular suite of statistical and machine learning techniques to calibrate and perform data corrections, with XGBoost being the default and most advanced option, and (iii) a post-processing module to calculate derivative variables and export the data in the self-describing and interoperable Measurement Data Format (MDF) (GEO Mountains, 2022; Fig. 9).

2.2　Data access and pre-processing

AtsMOS is currently designed to be used with data from the Global Forecasting System (GFS) from the US National Center for Atmospheric Research, which is freely available on a global scale and real-time basis (https://rda.ucar.edu/datasets/ds084.1/, last access: 28 October 2024). Alternative global or regional numerical prediction models, such as those produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) or national meteorological agencies, would also be suitable, particularly if real-time data can be accessed through an Application Programming Interface (API).

GFS forecasts are computed every 6 h, with a lead time from 0 to 384 h (16 d). Pressure-level data are generally preferred for mountain forecasting applications because the real-world surface in such regions is likely to be very different (e.g. in elevation and surface type) from the model surface (Mass et al., 2008), and hence we anticipate greater general predictability using data from the free atmosphere. We evaluated different methods for accessing GFS data and found that the web subsetting form is generally the most convenient for accessing historical archive data (http://rda.ucar.edu/datasets/ds084.1/dataaccess/, last access: 28 October 2024), while the online THREDDS server is best for downloading real-time data (https://tds.scigw.unidata.ucar.edu/thredds/catalog/idd/forecastModels.html, last access: 28 October 2024). As such we include a preset module in AtsMOS for both automatically downloading and pre-processing real-time data but only apply the pre-processing for the historical archive data. Historical archive data need only be downloaded once for pre-trained MOS models to be created (see below), which may then be run on any real-time data, ensuring the model remains computationally efficient.

We include example scripts in the AtsMOS Jupyter notebook used to pre-process instrumental data (Fig. 2). For instance, it is often necessary to synchronize measurement and NWP output timings – in the Everest example presented below instrumental data have a higher measurement frequency than the GFS data (6 h). We also include scripts for error checking and filtering of unreliable data, although we note that these are heavily dependent on the type and location of the sensor. We encourage users to carefully consider what, if any, processing steps are necessary for field data treated as "ground truth", as any errors or biases remaining will be learnt by the model. We discuss the limitations in further detail in the "Discussion and broader applicability" section.

2.3　Core machine learning

For the core processing, AtsMOS applies model output statistics (MOS) to the GFS data, with a range of possible correction algorithms for the user to select from depending on predictand type (e.g. binary or continuous) and the weighting of interpretability versus performance (Table 2).

In our case study below (Sect. 3), we compare the results from applying simple linear regression and XGBoost (Chen and Guestrin, 2016). Linear regression works well when the relationship between the predictor and target variable is approximately linear. Its coefficients provide clear insights into the impact of each feature, making it valuable for tasks where interpretability is crucial. However, linear regression cannot resolve nonlinear relationships in the data (without transformations to the input variables) and is sensitive to data quality and outliers limiting its predictive performance in many real-world cases. We implement a standard ordinary-least-squares-based linear regression algorithm in AtsMOS, which does not require any parameter choices.

XGBoost is at the other end of the complexity spectrum, combining decision trees with gradient boosting to improve computational efficiency and predictive performance, particularly in high-dimensional, nonlinear data scenarios. It has been shown to outperform most methods in terms of predictive accuracy (Chen and Guestrin, 2016) and is robust to overfitting, but its complexity can make it less suitable in cases where model transparency is essential and understanding the reasons behind incorrect predictions is key. A range of different parameters in XGBoost can be modified from their default values, including the type of objective function used (here, squared error), the learning rate (here, 0.1), the number of estimators (here, 250), the maximum tree depth (here, 4), and more (Chen and Guestrin, 2016). There is no parameter set that is optimal for all datasets. We tune the default parameters for AtsMOS based on the Mt. Everest case study described in Sect. 3, which we expect to be broadly applicable (if not optimal) for a wide range of cases. Users may easily run custom hyperparameter tuning for other custom datasets, but we do not include this in the default workflow due to its high computational cost.

AtsMOS is designed to be modular such that users can easily define new core machine learning (ML) processing algorithms where either methodological advances or specifics of their dataset demand a different approach. We implement linear regression, random forest, and XGBoost algorithms and note that most alternative ML techniques implemented in the scikit-learn Python package can also be used by changing a single line of code. An overview of the advantages and limitations of the different models implemented is provided in Table 2. Beyond the initial model evaluation (discussed in Sect. 3.3), we do not split the instrumental data into testing and training data. Instead, we train the ML model of choice using the full instrumental record to maximize both the volume and diversity of training data. We train a separate ML model for each forecast lead time as data error is expected to vary with lead time. All pre-trained ML models are saved once training is completed and can be directly loaded from file for future AtsMOS runs. In cases where no new instrumental data are available, this enables highly efficient runs in which model training can be bypassed entirely (Fig. 2). Where novel instrumental data are regularly available, we
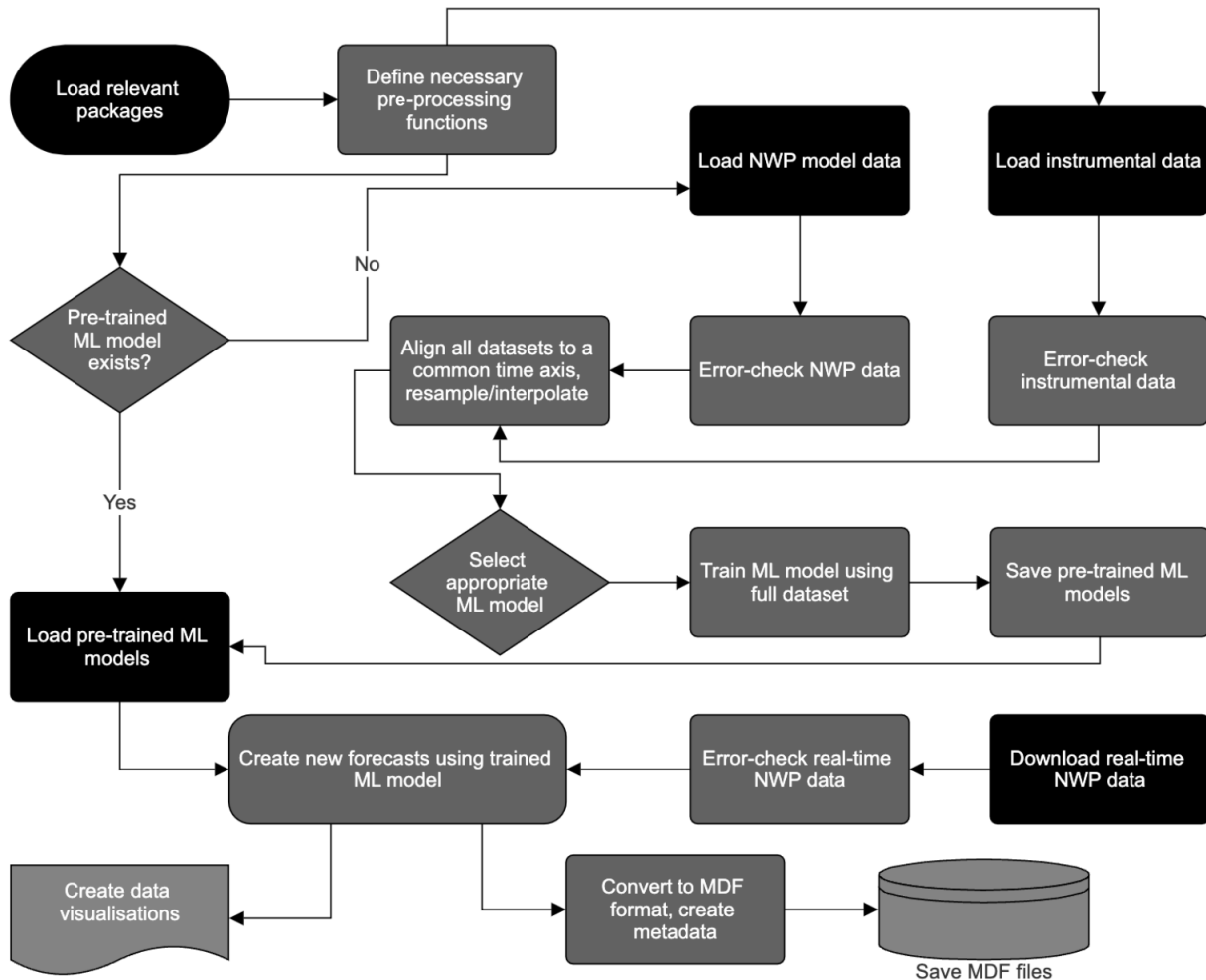
Figure 2. AtsMOS process workflow, described further in the documented AtsMOS Jupyter notebook at https://doi.org/10.5281/zenodo.10889509 (Van Wyk de Vries, 2024).

recommend periodic re-training of ML models to maximize expected forecast skill.

2.4 Post-processing and validation metrics

Once the appropriate ML model has been trained using the historical data, AtsMOS can process the real-time GFS forecast to produce corrected forecasts. The calibrated forecast may be a continuous variable (e.g. wind speed), a probability (e.g. probability of winds above a given threshold), or a binary categorized field (e.g. winds above or below a given threshold) depending on the processing choices made and project requirements. We also highlight that the flexible approach of AtsMOS enables prediction of any variable for which observations exist and which are sensitive to the atmospheric state. We showcase this in Sect. 3, making predictions for facial frostbite time – an important variable for mountaineering, which is not available as a direct output from any NWP model.

We use a range of possible metrics to evaluate model performance, including three primary metrics: Kling–Gupta efficiency (KGE; Gupta et al., 2009), mean absolute error (MAE), and root mean square error (RMSE). These metrics provide a comprehensive assessment of the model's accuracy, precision, and overall performance. KGE offers a balanced evaluation by combining correlation, bias, and variability, making it particularly suitable for hydrological and meteorological applications. MAE measures the average magnitude of errors, providing a straightforward interpretation of forecast accuracy. RMSE, on the other hand, emphasizes larger errors, highlighting potential issues in model predictions. Additionally, our implementation supports a variety of other metrics which may be added if users require, including R^2, residual skewness, residual kurtosis, slope, intercept, Nash–Sutcliffe efficiency (NSE), correlation, relative variance, and bias, allowing for a thorough validation of the model across different aspects of performance and different expected error profiles.

Table 2. Different machine learning algorithms currently implemented in AtsMOS, along with strengths and weaknesses. AtsMOS is modular, and users can easily define new core processing algorithms.

Algorithm	Description	Strengths	Weaknesses
Linear regression	Regression using ordinary least-squares fit	Computationally efficient, explainable	Sensitive to outliers, liable to over-fitting, many assumptions (linearity, normality of errors), cannot resolve nonlinear relationships in the data
Random forest	Scikit-learn ensemble decision tree regressor	Computationally efficient for large parameter spaces, robust to multiple non-dependent variables, allows for easy inspection of feature importance to enhance model interpretability	Computationally expensive for large datasets, lower reliability for unbalanced datasets, more complex than regression
XGBoost	Optimized distributed gradient boosting and parallel tree boosting algorithm	Most advantages listed for random forest, computationally efficient for large parameter spaces, highest accuracy method in several machine learning competitions	Limited interpretability relative to other algorithms, remains liable to overfitting for small training datasets, cannot reasonably extrapolate beyond the range captured in the training data

2.5 Data export

As a final stage in AtsMOS, the corrected forecast variables are saved along with their metadata in the self-describing and interoperable MDF (GEO Mountains, 2022). This final export stage has the benefit of (i) enabling easy usage of the custom forecasts in other applications, or plotting dashboards; (ii) ensuring that the variables are saved with all necessary context for long-term archiving; and (iii) through standardized nomenclature, enabling easy comparison with other forecast datasets and external validation.

Overall, the AtsMOS workflow is designed to be lightweight and flexible, while enhancing the predictive skill of large-scale forecasts using local observations.

3 Example application: Mt. Everest summit meteorology

3.1 Background

As the highest peak on Earth, Mt. Everest sees hundreds of attempts to scale its 8850 m a.s.l. summit each year. Fatalities are common, including 17 fatalities in spring 2023 (Ellis-Petersen, 2023) and an overall mortality rate of around 1 % over the past decade (Huey et al., 2020). Weather is a major contributor to these, playing a role in 25 % of deaths (Firth et al., 2008) due to the hazard from extremely low barometric pressure (low oxygen availability) and severe cold that climbers may be exposed to (Moore and Semple, 2006; Matthews et al., 2020a, 2022). The latter is sensitive to wind speed (Moore and Semple, 2011), which if high enough may also directly blow climbers off the mountain. There-

fore, climbers limit their summit attempts to periods when the subtropical jet's retreat leaves lower wind speeds on the mountain. Therefore, accurately forecasting these periods of lighter winds is critical for minimizing the risk of climbing Mt. Everest.

Whilst deciding the acceptable wind speed threshold for summit attempts is dependent on individual climbers' risk tolerance, physical considerations suggest that a human with an effective surface area (A_p) of $0.5\,\mathrm{m}^2$ is at risk of being blown over if the wind force (F) exceeds 72 N (Hugenholtz and VanVeller, 2016; McIlveen, 2002). F is related to the wind speed (v) according to

$$F = \frac{1}{2}\rho v^2 A_p C_D, \tag{1}$$

where ρ is the air density ($\mathrm{kg\,m}^{-3}$) and C_D is the drag coefficient (dimensionless). Using $C_D = 0.6$ from McIlveen (2002), the critical wind speed (v_c) yielding 72 N can be evaluated:

$$v_c = \sqrt{\frac{144}{(0.3\rho)}}. \tag{2}$$

At the altitude of the highest camp (the South Col: 7945 m a.s.l.; Fig. 3) on Mt. Everest's main route from Nepal – which marks the beginning of the "death zone" – ρ (which depends on temperature and pressure) is, on average, $0.52\,\mathrm{kg\,m}^{-3}$, translating to $v_c = 30.3\,\mathrm{m\,s}^{-1}$ according to data from May 2019 until June 2023 (see the following section). To illustrate the use of AtsMOS for the delivery of decision-critical forecasts we therefore use a new network of Mt. Everest weather stations (see below) to develop predictions of (1) absolute wind speed and (2) the probability of
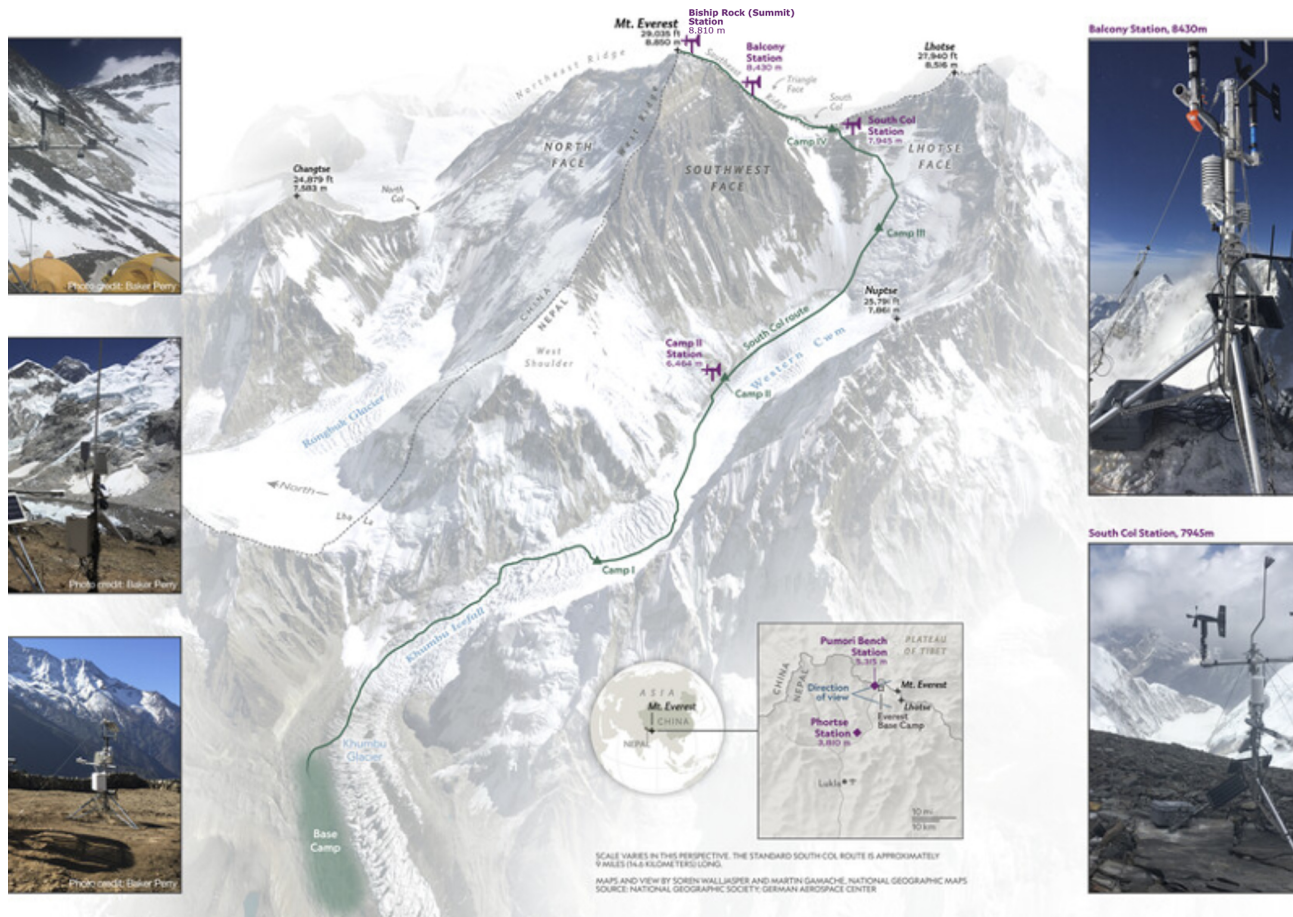
Figure 3. Location of the weather stations on Mt. Everest. Modified after Matthews et al. (2020b).

speeds exceeding both 30 and $20\,\mathrm{m\,s^{-1}}$. The upper threshold is used to identify dangerous winds (high hazard), whilst the lower we regard as potentially dangerous (medium hazard) and hence a conservative threshold for identifying suitable weather for a summit attempt. We also showcase the flexibility of AtsMOS to directly forecast key variables such as wind chill temperature and facial frostbite time.

3.2 Mount Everest weather data

In spring 2019, a network of five automated weather stations was installed on the Nepali side of Everest, known locally as Sagarmatha or Qomolangma, including three stations above the base camp at Camp 2 (6464 m), the South Col (7945 m), and the Balcony (8430 m; Matthews et al., 2020a). Of these, the two highest stations, the South Col (7945 m a.s.l.) and the Balcony (8430 m a.s.l.), were positioned to monitor the potentially dangerous winds on the upper mountain. However, the Balcony's record is relatively short (due to wind damage) and considered unrepresentative of the upper mountain due to sheltering under common flow directions. A further station was installed at 8810 m altitude on the highest-elevation

exposed bedrock near the summit (the "Bishop Rock") in spring 2022, which is currently the highest-altitude weather station in the world with publicly available data (Matthews et al., 2022). Note that another station was installed by a Chinese team at a similar altitude on the north side of the mountain in 2022, although its status and data availability are unknown.

Three of the four weather stations (Bishop Rock or "Summit", Balcony, and South Col) were installed with two separate wind speed sensors. The dual sensors were installed for redundancy in the event of one failing but are also valuable for evaluating the reliability of wind speed observations. Recovery of destroyed monitoring equipment showed that the wind speed sensors can suffer from mechanical failure (breakage of the anemometer cups) and growth of rime ice that result in incorrect measurements, but that is not evident from a single time series. Therefore, we apply a moving window cross-correlation between the two sensors' time series in the pre-processing stage of AtsMOS to identify periods of decorrelation and unreliable data. We use a minimum correlation threshold of 0.9 measured over a 14 d window for both the mean and maximum hourly wind speed (measured
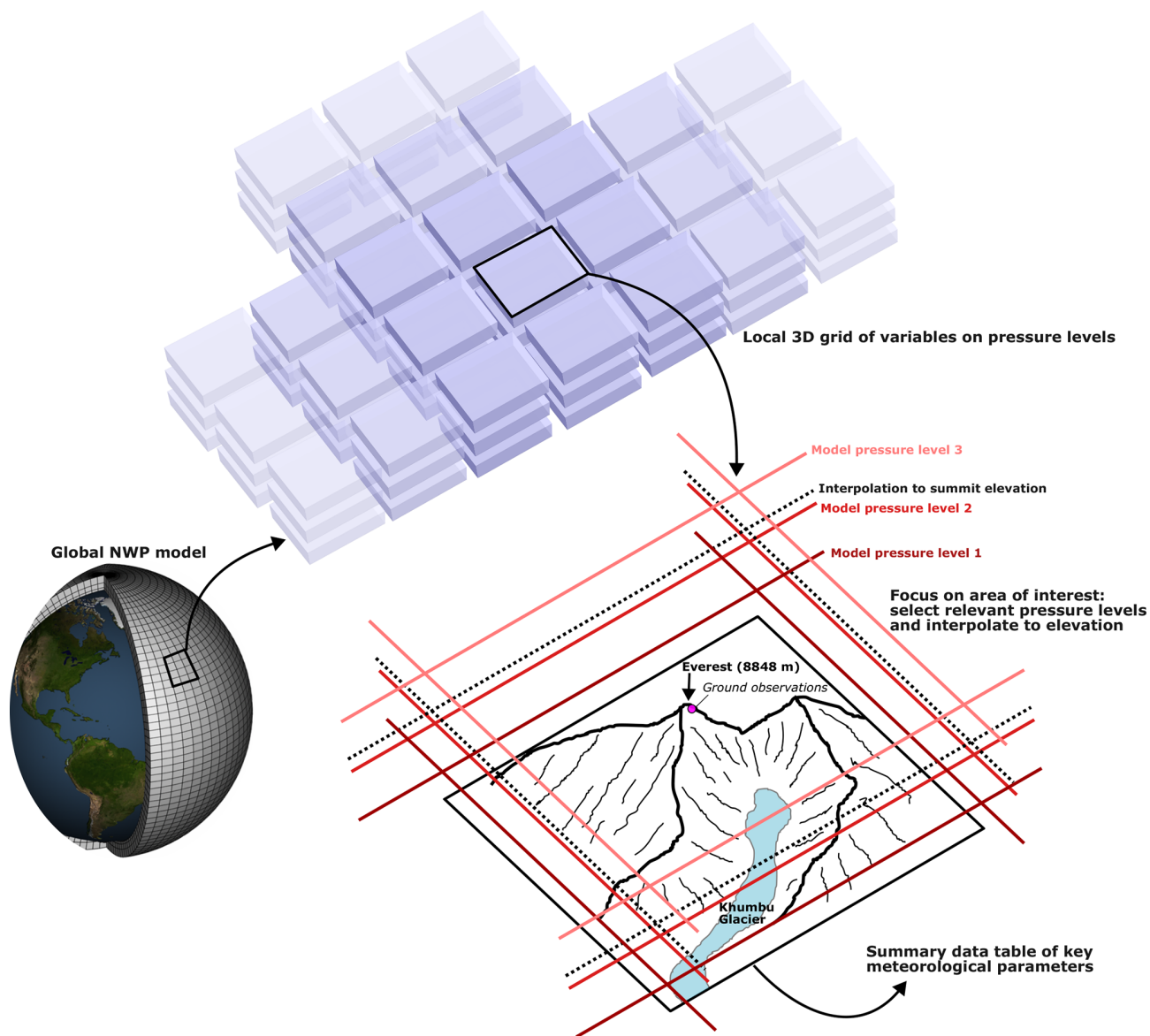
Figure 4. Module for extracting key meteorological parameters from the global numerical weather prediction (NWP) model around a point of interest (typically, the location of the ground observations). Both horizontal and vertical derivatives are calculated from the NWP data to supplement the ML training dataset.

at 5 s intervals) to determine reliable data and mask out data points falling below this threshold (Fig. 5, top two rows).

Only the South Col station has a data record covering a period longer than a few months and across multiple years. While this station is located almost 1000 m below the summit, its position at the head of the Khumbu Valley with an open westerly aspect (the dominant wind direction) means that its wind speeds are very similar to the summit (Pearson's r value $= 0.85$). We apply the dual-sensor correlation threshold (0.9) and filter out winds with a direction outside the range $270 \pm 45°$ due to the risk of topographic shielding outside this window. The lower elevation leads to a slight neg-

ative bias in wind speed at the South Col, which is on average 18 % slower than summit winds. We linearly regress the remaining South Col wind record against the filtered summit record and use this to create a synthetic summit record. The resulting record contains just over 1 year of data, spread across two 6-month periods from June 2019 to January 2020 and May 2022 to January 2023 (Fig. 5).

For the NWP component of the AtsMOS loading and pre-processing stage, data were loaded and pre-processed from the Global Forecasting System (GFS) (https://rda.ucar.edu/datasets/ds084.1/, last access: 28 October 2024) We downloaded all 10 variables: precipitation, temperature, relative
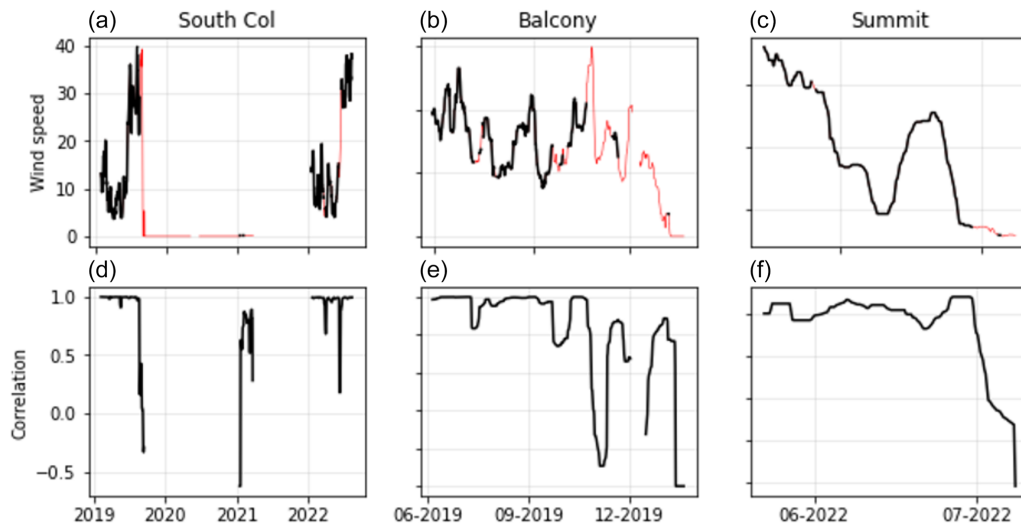
Figure 5. Validation of observational time series. Wind speed data shown in black **(a–c)** were considered reliable based on the dual-sensor correlation **(d–f)**, while wind speed data show in red were judged unreliable and removed.

humidity, N–S wind speed, E–W wind speed, vertical velocity, geopotential height, absolute vorticity, cloud mixing ratio, and ozone mixing ratio. We choose to include all variables (irrespective of whether physical connections to the predictand could be identified a priori) because (i) their variations could provide insight into relevant sub-grid-scale process, and (ii) the default machine learning method we select (XGBoost) is robust to overfitting and collinearity. A user-supplied list of variable names can also be supplied to AtsMOS to limit the variables used in model fitting. The data were downloaded for a $9 \times 9 \times 3$ data cube centred on the summit of Sagarmatha/Qomolangma, with nine data points in each horizontal direction (from 27–29° latitude and 86–88° longitude at 0.25° spacing) and three vertical pressure levels (350, 400, and 450 hPa). We use the geopotential height from the three pressure levels to linearly interpolate or extrapolate all variables to a fourth vertical level, corresponding to the summit elevation at 8849 m. Finally, we calculate the horizontal and vertical gradients in the nine variables to further account for potential drivers of relevant sub-grid-scale processes. A full list of all 172 variables and derivatives used is in the Supplement. We separately download the GFS historical archive (via the NCAR web portal) and real-time GFS forecast (programmatically from the THREDDS server – see notebook), with the former used to calibrate our data corrections and the latter used to produce corrected forecasts.

For the core processing component of AtsMOS we use (simple) linear regression and (complex) XGBoost algorithms to improve the GFS forecast for the wind speed at the summit of Sagarmatha/Qomolangma. To avoid issues with temporal autocorrelation of training and validation data, we split our time series in half in January 2021. This test–train split provides us with 6 months of training data and 6 months

of validation data from June 2019 to January 2020 and May 2022 to January 2023. We run each MOS algorithm twice: once training on data from 2019 and testing on data from 2022 and vice versa. Linear regression is applied using just the GFS model wind speed interpolated to the 8849 m summit altitude as the only predictor variable; XGBoost, on the other hand, is trained using all 172 GFS variables and spatial derivatives. We reproduce the simple and complex MOS workflows for several GFS lead times: analysis (0 nowcast), 24 h, 48 h, 120 h (5 d), and 240 h (10 d).

While predicting Sagarmatha/Qomolangma wind speed as a continuous variable is scientifically interesting, a categorical prediction of dangerous versus safe winds may be of more use to the majority of potential end users (sherpas and mountaineers). We therefore employ a wind speed threshold of $30\,\mathrm{m\,s^{-1}}$ to classify our synthetic wind time series into a time series of dangerous winds. We also use a second lower threshold of $20\,\mathrm{m\,s^{-1}}$ to classify potentially dangerous winds. A wind speed of $30\,\mathrm{m\,s^{-1}}$ approximately corresponds to the wind speed required to blow a human off their feet at Sagarmatha/Qomolangma summit conditions (Sect. 3.1). We intentionally do not call winds below this threshold "safe" as they can still be hazardous in a variety of ways (including slowing ascents and increasing exposure), but they correspond to conditions during which – at least in principle – a typical climber should not be in danger of being blown from the mountain. We use the same XGBoost MOS to run the categorical forecast using GFS lead times of 0 h (analysis), 48 h, and 240 h (10 d). For the 0 and 48 h lead times we forecast dangerous winds at the native GFS 6 h temporal resolution. For the 240 h (10 d) lead time, however, we invert the problem and classify 48 h (2 d) periods during which all winds are below the given threshold. The objective of classifying low-wind periods with a 10 d lead time is to enable earlier

identification of favourable summit weather conditions and a better distribution of climbs throughout the season to prevent potentially dangerous overcrowding.

3.3 Results and model evaluation

We test the AtsMOS dataset by training it on data from the first period (2019–2020) and predicting data over the second period (2021–2022) and vice versa. This enables a more robust validation than random test–train splitting of the dataset by reducing inflation of model ability caused by meteorological time series temporal autocorrelation. We evaluate three different learning techniques: simple linear regression, random forest, and XGBoost (Figs. 6–8).

Linear regression produces a reasonable overall fit to the data (Fig. 6), with a model wind speed–field-measured wind speed R^2 of 0.87, a root mean squared error (RMSE) of $10.59 \, \mathrm{m\,s^{-1}}$, and a mean absolute error (MAE) of $7.87 \, \mathrm{m\,s^{-1}}$. The Kling–Gupta efficiency of these datasets is 0.73, evaluating a combination of their correlation, relative variation, and mean bias and with higher values reflecting a better fit (Gupta et al., 2009). In particular, linear regression successfully matches the magnitude of winds during the majority of the low-wind (monsoon) season from July to October. However, it fails to match the magnitude of the highest-wind-speed events, with a clear overestimate evident.

Random forest produces a good overall fit to the data, with a model wind speed–field-measured wind speed R^2 of 0.92 and a root mean squared error (RMSE) of $8.52 \, \mathrm{m\,s^{-1}}$, and a mean absolute error (MAE) of $6.33 \, \mathrm{m\,s^{-1}}$. The Kling–Gupta efficiency of these datasets is 0.77. There are three notable improvements of the model trained with random forest regression relative to standard linear regression: the estimates are more closely clustered along the 1 : 1 model–data line, the timings of high-wind episodes in the model better match those observed in the data, and the magnitude of high-wind peaks matches better across both datasets – although a small bias towards higher winds than reality remains.

XGBoost produces a good fit to the data, with a model wind speed–field-measured wind speed R^2 of 0.93, a root mean squared error (RMSE) of $7.95 \, \mathrm{m\,s^{-1}}$, and a mean absolute error (MAE) of $5.97 \, \mathrm{m\,s^{-1}}$. The Kling–Gupta efficiency of these datasets is 0.79. The overall performance of the model trained with XGBoost is similar to that trained with random forest, with a slightly improved fit across all metrics. The timing of high-wind events is well predicted, and while the model still tends to overestimate the magnitude of high-wind events, the bias is lower (bias score: 0.86 for XGBoost relative to 0.84 for random forest and 0.81 for linear regression).

We then apply the AtsMOS workflow to a real-time case study for the approximately 2-week (384 h) period from 20 July 2023 to 5 September 2023 using GFS forecast data as described in the methods. As well as calculating the wind speed, temperature, and precipitation, we compute forecasts of two derivative variables: wind chill temperature and facial frostbite time (Moore and Semple, 2011). Both wind chill temperature and facial frostbite time are calculated based on wind speed and temperature forecasts according to the formulas of Moore and Semple (2011). Wind chill temperature reaches as low as $-45\,^{\circ}\mathrm{C}$ on 3 September 2023, also aligning with the shortest facial frostbite time of less than 7 min (Fig. 9). Forecast wind speeds do not exceed $20\,\mathrm{m\,s^{-1}}$ but reach more than $19\,\mathrm{m\,s^{-1}}$ on the night of 20–21 July 2023, with the short forecast lead reducing the uncertainty for the forecast. The facial frostbite time briefly falls below 10 min on this night, also driven by the high wind speeds, and the wind chill temperature fluctuates between -35 and $-40\,^{\circ}\mathrm{C}$ – well below the air temperature ($-20\,^{\circ}\mathrm{C}$), highlighting the importance of the wind speed in modulating the cold hazard and thereby the value of computing this derived variable with AtsMOS.

4 Discussion and broader applicability

The AtsMOS workflow builds on advancements in machine learning and data accessibility to improve mountain weather forecasts by downscaling coarse numerical model outputs to specific locations of high value. Through a case study focusing on Mt. Everest summit meteorology, we demonstrate the effectiveness of AtsMOS in refining wind speed (and wind chill temperatures) critical for assessing risks for mountaineering. This workflow is open-source, flexible, and computationally cheap – enabling more accurate mountain weather predictions across many different environments.

Our Mt. Everest case study showcases a local application of the AtsMOS workflow for an environment in which mountaineers and sherpas knowingly expose themselves to potentially deadly conditions (Moore and Semple, 2006, 2011; Matthews et al., 2020b). Therefore, more precise meteorological forecasts are critical for expedition planning. First, by more accurately predicting wind speeds, our system enables expedition organizers to identify windows of potentially safe (lower wind) conditions with approximately 2 weeks' notice, allowing the timing of trips to the upper mountain to be determined at an earlier date. This is invaluable for optimizing expedition scheduling, maximizing the likelihood of successful summit attempts, and potentially improving safety by preventing dangerous overcrowding from teams rushing to exploit weather windows at late notice. Second, the shorter lead time and more precise AtsMOS forecasts assist in preventing climbs during times of dangerous weather, thereby enhancing safety. By providing reliable forecasts for both dangerous and potentially dangerous wind thresholds, our workflow empowers expedition leaders to make informed decisions, avoiding ascent attempts during periods of heightened risk.

The flexible nature of our workflow enables outputs with different levels of complexity (e.g. Fig. 9), ranging from binary classifications ("dangerous/potentially safe") and raw
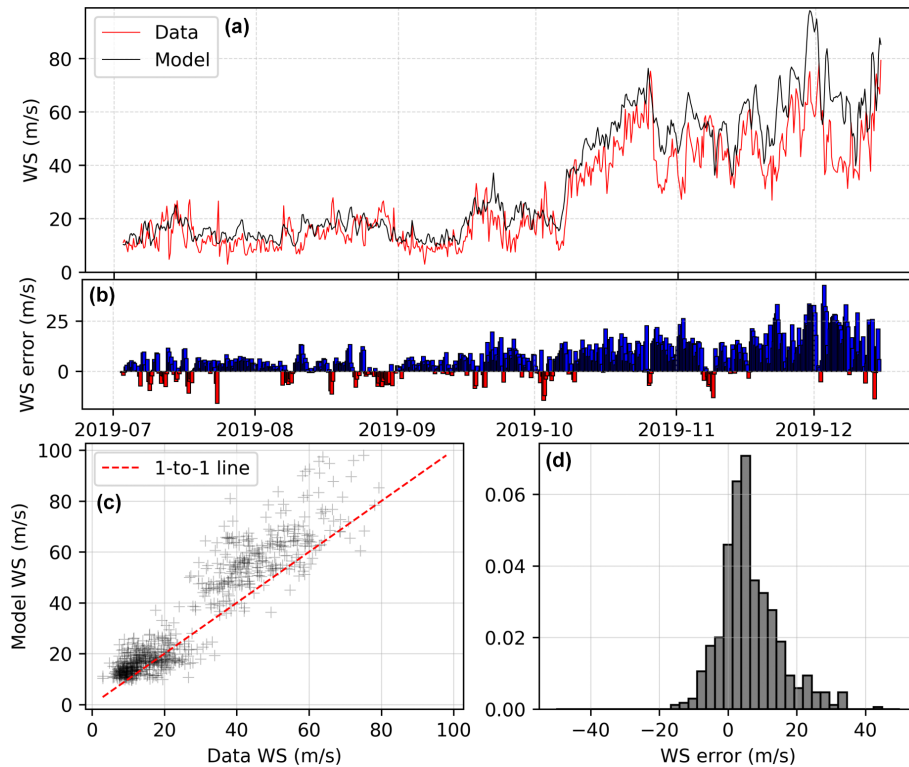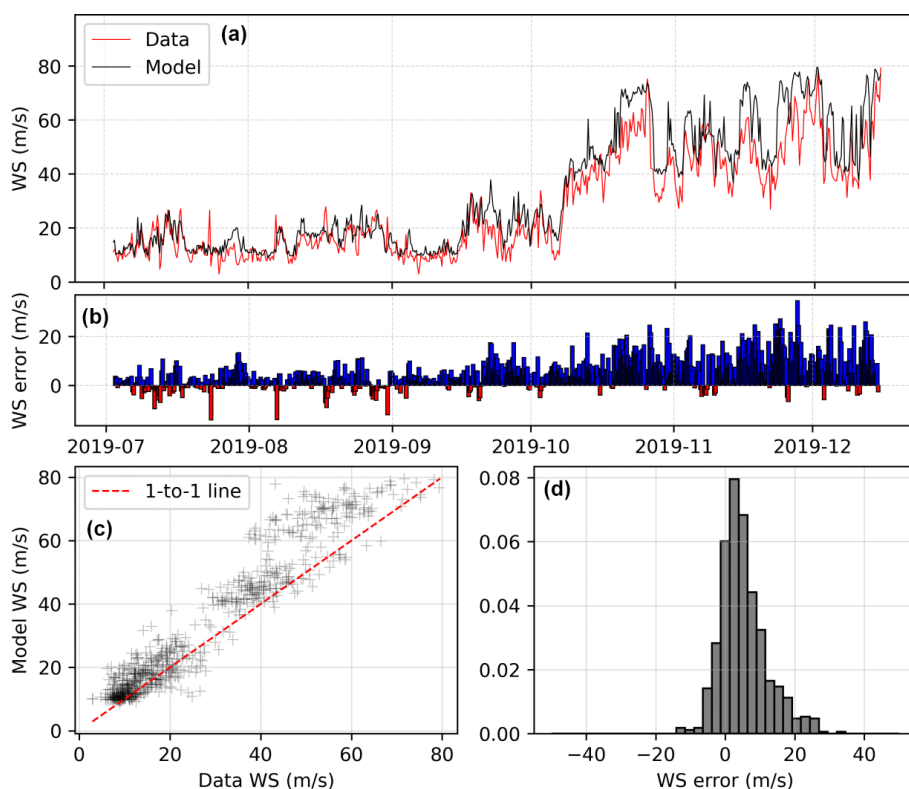
Figure 6. (a) Observed (red) and modelled (black) wind speed for the first observational period at the summit, with model (here, linear regression) training using only the second period (2019). **(b–d)** Difference between modelled and observed wind speed shown as a difference bar chart, scatterplot, and histogram, respectively. The statistics are as follows. $R^2 = 0.87$, root mean squared error (RMSE): 10.59, skew: 0.37, kurtosis: 1.25, mean absolute error (MAE): 7.87, Nash–Sutcliffe efficiency (NSE): 0.78, Kling–Gupta efficiency (KGE): 0.73, correlation: 0.93, relative variance: 0.81, bias: 0.81.

meteorological variable forecasts (wind speed, temperature, etc.) to derivative variables (e.g. facial frostbite time). This flexibility offers a wide range of possibilities to enable expedition planners, who, armed with more information, should be able to plan safer climbs, thereby reducing the risk of attempting to climb this iconic mountain. One example is shown in Fig. 9c, with "medium" and "high" hazard times delineating periods with high probabilities of strong winds. The evaluation of hazard probability with AtsMOS is seen as a particularly important feature for end users. If properly calibrated, it more clearly aligns the forecast product with decision-making. Without the MOS approach here, expedition planners would likely need to consult ensemble forecasts (e.g. the Global Ensemble Forecasting System) to produce comparable probabilities, associated with a non-trivial increase in data processing for support teams and/or the burden on the expedition planner to interpret the forecast. Of course, we also note here that for AtsMOS to be used within such ensemble forecasting systems, more steps may be required – for example, propagating the ensemble members through the ML algorithms calibrated on the deterministic forecast to explore uncertainty more fully. This can ultimately combine the

benefits of both the ensemble forecast and reduced bias from local calibration.

While the AtsMOS workflow's potential to improve local mountain meteorology forecasts is promising, it is important to acknowledge its limitations. The most significant constraint lies in the workflow's dependence on the two underlying data sources: numerical weather prediction data and instrumental and/or field data. AtsMOS outputs therefore rely on the assumption that, while these datasets may contain uncertainties or noise, they both contain real and useful information about local meteorological conditions. There are a number of scenarios in which this may not be the case for either dataset: for instance, large-scale NWP models missing key local processes (leading, for example, to a poor representation of convection) or sensors becoming degraded and recording false data (for instance, a wind sensor covered in rime ice). This limitation is present at both the training and prediction stages of the process. The effectiveness of the workflow is, therefore, highly dependent upon the availability and quality of ground observations, which are particularly rare in remote and high-altitude regions like Mt. Everest (Matthews et al., 2020a; Thornton et al., 2022). The applicability of AtsMOS may also be limited in regions with
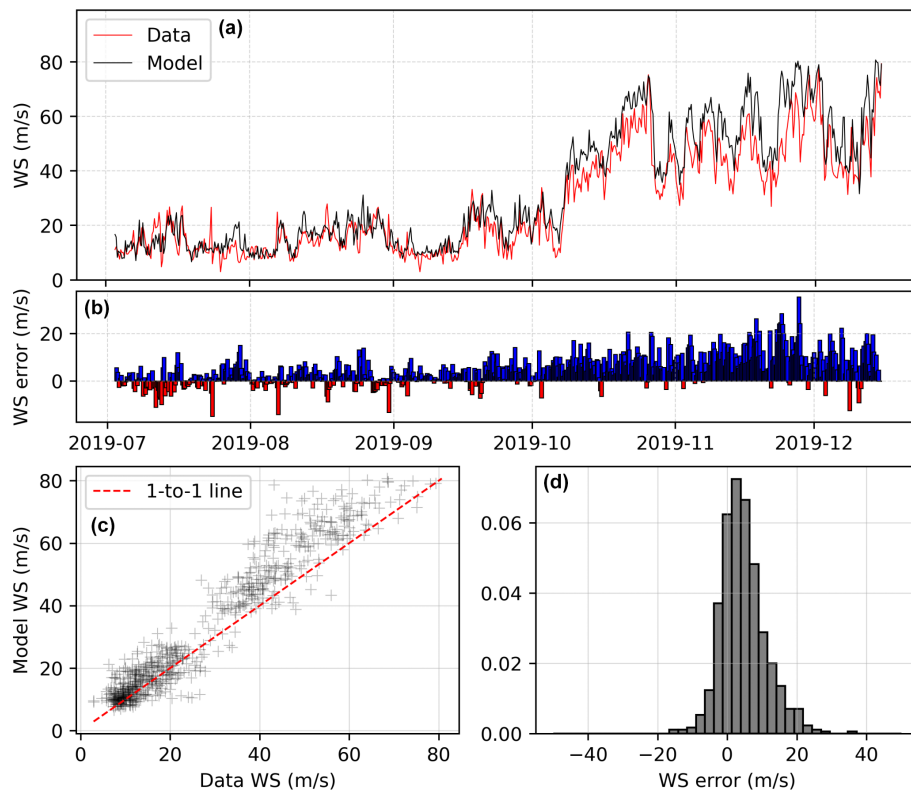
Figure 7. (a) Observed (red) and modelled (black) wind speed for the first observational period at the summit, with model (here, random forest) training using only the second period (2019). **(b–d)** Difference between modelled and observed wind speed shown as a difference bar chart, scatterplot, and histogram, respectively. The statistics are as follows. $R^2 = 0.92$, RMSE: 8.52, skew: 0.39, kurtosis: 1.2, MAE: 6.33, NSE: 0.84, KGE: 0.77, correlation: 0.96, relative variance: 0.84, bias: 0.84.

unique or extreme meteorological conditions not adequately captured by existing NWP models, even with the aid of machine learning to extract additional information, which may be of concern if these regions are of particular interest for hazard mitigation. We note, however, that the latter may be guarded against by using near-real-time (i.e. lagged) observations from the telemetry-enabled weather stations (Chkeir et al., 2023).

We also highlight the need for caution in the application of machine learning algorithms. Whilst techniques like random forest and XGBoost can offer enhanced predictive capabilities, they may also introduce complexities in model interpretation and require careful validation to ensure robust performance. These limitations underscore the need for ongoing refinement and validation of the workflow to optimize its utility and effectiveness in diverse mountainous environments. One specific concern in the usage of tree-based machine learning algorithms such as random forest or XGBoost is that they cannot reasonably extrapolate beyond the range captured in the training data. This is a particular concern in areas with strong seasonal variation, where training on one season alone may lead to failure to produce meaningful predictions in the other season. In the case of Everest, this limitation is mitigated by having data covering the transition from the low-

to high-wind season, but in areas where this is not possible alternative methods may need to be considered.

Another type of "overfitting" may occur if machine learning inadvertently reproduces biases in the observations, for example, due to instrumentation errors. This challenge should be taken seriously, as the error could be systematic and dangerous. For example, if icing of wind sensors occurred preferentially in conditions of low temperature and high winds (i.e. periods of greater cold stress), the machine learning, trained on the errors, would underestimate the hazard most when it was greatest. Such risks highlight the importance of thoroughly quality-checking the observations in the pre-processing stage of the AtsMOS workflow. We note that, on Mt. Everest, the station design enables the detection of such icing through the use of redundant wind sensors (Matthews et al., 2020a, 2022). We hope that ongoing efforts to develop a universal high-altitude observing platform (to enhance mountain weather monitoring worldwide) will also be designed with such challenges in mind (Napoli et al., 2023). More generally, we emphasize that the AtsMOS approach to forecast improvement differs from efforts to embed ML in NWP (e.g. Frnda et al., 2022). In this case, ML algorithms do not replace high-quality observational data; rather they emphasize the need for it and amplify any limitations of

Figure 8. (a) Observed (red) and modelled (black) wind speed for the first observational period at the summit, with model (here, XGBoost) training using only the second period (2019). **(b–f)** Difference between modelled and observed wind speed shown as a difference bar chart, scatterplot, and histogram, respectively. The statistics are as follows. $R^2 = 0.93$, RMSE: 7.95, skew: 0.29, kurtosis: 1.69, MAE: 5.97, NSE: 0.86, KGE: 0.79, correlation: 0.96, relative variance: 0.85, bias: 0.86.

the data. By investing in data quality and instrumentation and leveraging ML alongside this, we increase our potential for accurate and actionable meteorological forecasts in mountainous regions.

In addition to ensuring the accuracy and reliability of sensor data, effective data management practices are crucial for maximizing the utility and impact of field datasets, particularly in the context of mountain meteorology. Good metadata, which provide detailed information about the characteristics and origins of the data, is essential for understanding and interpreting observational datasets. Interoperability, where data can be integrated and exchanged across different platforms and systems with minimal barriers, becomes increasingly important when considering the generalizability of findings and methodologies to other mountainous environments. While the specific challenges and characteristics of each mountain region may vary, the fundamental principles and approaches developed for mountain meteorology research can often be applied more broadly and insights and techniques developed in one region can inform and benefit studies in others. Promoting robust data management practices is key for both the effectiveness of individual research efforts and the broader advancement of mountain meteorology as a field.

Whilst we have demonstrated the added value of improving weather forecasts for Mt. Everest with AtsMOS, we anticipate much greater benefits from this approach than just improving the safety of mountaineering expeditions. For instance, the ability to forecast thresholds for rainfall-triggered landslides, snow avalanches, or flooding relies heavily on accurate meteorological data and predictive models. By integrating high-resolution local meteorological data from AtsMOS into early warning systems, communities can better prepare for and respond to extreme weather events, reducing the risk of casualties and damage. Furthermore, on a regional or national scale, the integration of detailed mountain meteorology datasets into larger-scale networks enhances the effectiveness of early warning systems by providing comprehensive coverage of weather patterns and potential hazards across diverse landscapes. Improved prediction of meteorological conditions in mountainous regions has far-reaching implications for promoting the resilience and safety of mountain communities and ecosystems and is an important component of effective early warning systems for many hazards.
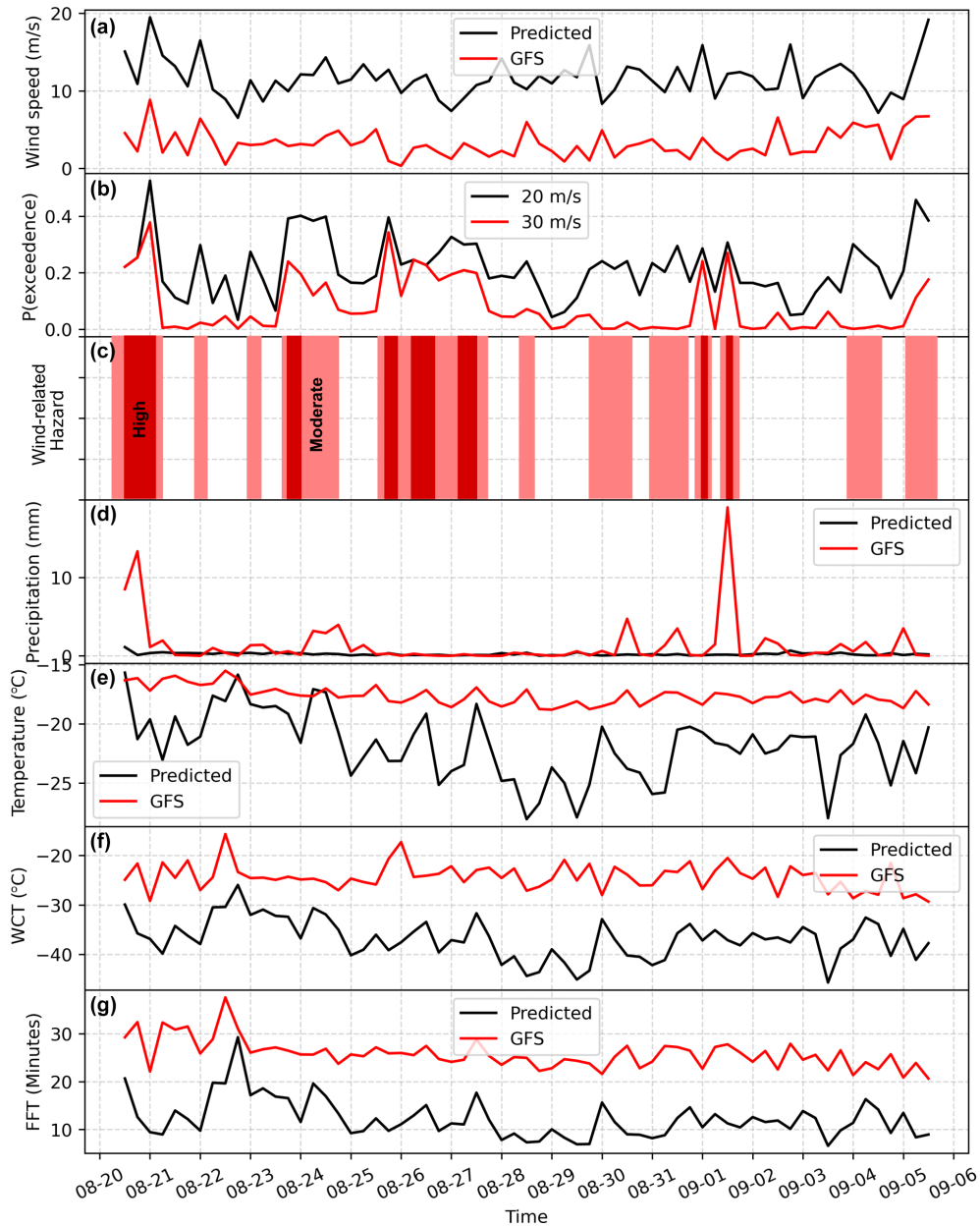
Figure 9. Example of a real-time forecast for wind (**a**), precipitation (**d**), temperature (**e**), and derivate variables of wind chill temperature (**f**, WCT) and facial frostbite time (**g**, FFT; Moore and Semple, 2011). The raw GFS estimates of wind speed and precipitation are shown in red alongside our predictions; note how the AtsMOS wind speed estimates are substantially higher and precipitation is substantially higher. The probability of exceedance of different wind speed thresholds, 20 and $30\,\mathrm{m\,s^{-1}}$, is also shown in (**b**) with related hazard classification, with medium hazard denoted as $> 20\,\%$ probability of winds exceeding $20\,\mathrm{m\,s^{-1}}$ and high hazard as $> 20\,\%$ probability of winds exceeding $30\,\mathrm{m\,s^{-1}}$.

5 Conclusion

In conclusion, the AtsMOS workflow represents a computationally efficient template for downscaling numerical model outputs using one observation or a small number of field observations. The template outlines a flexible, modular workflow for custom pre-processing of field observations or numerical weather model outputs depending on the need and provides several possible core learning algorithms ranging from simple linear regression to more complex random forest and XGBoost. We explore an example application at Mt. Everest, which demonstrates its practical utility in improving the prediction of critical weather parameters for mountaineering safety. There are limitations to this approach, including reliance on high-quality sensor data and potential

biases inheritance in machine learning algorithms. Moving forward, continued research and observation network development hold promise for improving the accuracy and reliability of mountain meteorology forecasts, ultimately enhancing hazard mitigation efforts, and contributing to the resilience of communities living in these landscapes.

References

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, https://doi.org/10.1038/nature14956, 2015.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Chkeir, S., Anesiadou, A., Mascitelli, A., and Biondi, R.: Nowcasting extreme rain and extreme wind speed with machine learning techniques applied to different input datasets, Atmos. Res., 282, 106548, https://doi.org/10.1016/j.atmosres.2022.106548, 2023.

Corbari, C., Ravazzani, G., Perotto, A., Lanzingher, G., Lombardi, G., Quadrio, M., Mancini, M., and Salerno, R.: Weekly Monitoring and Forecasting of Hydropower Production Coupling Meteo-Hydrological Modeling with Ground and Satellite Data in the Italian Alps, Hydrology, 9, 29, https://doi.org/10.3390/hydrology9020029, 2022.

Ellis-Petersen, H.: Climate change to blame for up to 17 deaths on Mount Everest, experts say, The Guardian, https://www.theguardian.com/world/2023/may/30/climate-change-to-blame-for-up-to-17-deaths-on-mount-everest-experts-say (last access: 28 October 2024), 2023.

Firth, P. G., Zheng, H., Windsor, J. S., Sutherland, A. I., Imray, C. H., Moore, G. W. K., Semple, J. L., Roach, R. C., and Salisbury, R. A.: Mortality on Mount Everest, 1921–2006: descriptive study, BMJ, 337, a2654, https://doi.org/10.1136/bmj.a2654, 2008.

Frnda, J., Durica, M., Rozhon, J., Vojtekova, M., Nedoma, J., and Martinek, R.: ECMWF short-term prediction accuracy improvement by deep learning, Sci. Rep.-UK, 12, 7898, https://doi.org/10.1038/s41598-022-11936-9, 2022.

Glahn, B.: Determining an Optimal Decay Factor for Bias-Correcting MOS Temperature and Dewpoint Forecasts, Weather Forecast., 29, 1076–1090, https://doi.org/10.1175/WAF-D-13-00123.1, 2014.

Glahn, H. R. and Lowry, D. A.: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting, J. Appl. Meteorol. Clim., 11, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2, 1972.

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., and Hoefler, T.: Deep learning for post-processing ensemble weather forecasts, Philos. T. R. Soc. A, 379, 20200092, https://doi.org/10.1098/rsta.2020.0092, 2021.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Han, L., Chen, M., Chen, K., Chen, H., Zhang, Y., Lu, B., Song, L., and Qin, R.: A Deep Learning Method for Bias Correction of ECMWF 24–240 h Forecasts, Adv. Atmos. Sci., 38, 1444–1459, https://doi.org/10.1007/s00376-021-0215-y, 2021.

Herman, G. R. and Schumacher, R. S.: "Dendrology" in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us about Forecasting Extreme Precipitation, Mon. Weather Rev., 146, 1785–1812, https://doi.org/10.1175/MWR-D-17-0307.1, 2018.

Huey, R. B., Carroll, C., Salisbury, R., and Wang, J.-L.: Mountaineers on Mount Everest: Effects of age, sex, experience, and crowding on rates of success and death, PLOS ONE, 15, e0236919, https://doi.org/10.1371/journal.pone.0236919, 2020.

Hugenholtz, C. H. and VanVeller, G. S.: Wind hazard in the alpine zone: a case study in Alberta, Canada, Weather, 71, 27–31, https://doi.org/10.1002/wea.2567, 2016.

Immerzeel, W. W., Lutz, A. F., Andrade, M., Bahl, A., Biemans, H., Bolch, T., Hyde, S., Brumby, S., Davies, B. J., Elmore, A. C., Emmer, A., Feng, M., Fernández, A., Haritashya, U., Kargel, J. S., Koppes, M., Kraaijenbrink, P. D. A., Kulkarni, A. V., Mayewski, P. A., Nepal, S., Pacheco, P., Painter, T. H., Pellicciotti, F., Rajaram, H., Rupper, S., Sinisalo, A., Shrestha, A. B., Viviroli, D., Wada, Y., Xiao, C., Yao, T., and Baillie, J. E. M.: Importance and vulnerability of the world's water towers, Nature, 577, 364–369, https://doi.org/10.1038/s41586-019-1822-y, 2020.

Khadka, A., Wagnon, P., Brun, F., Shrestha, D., Lejeune, Y., and Arnaud, Y.: Evaluation of ERA5-Land and HARv2 Reanalysis Data at High Elevation in the Upper Dudh Koshi Basin (Everest Region, Nepal), J. Appl. Meteorol. Clim., 61, 931–954, https://doi.org/10.1175/JAMC-D-21-0091.1, 2022.

Lagerquist, R., McGovern, A., and Smith, T.: Machine Learning for Real-Time Prediction of Damaging Straight-Line Convective Wind, Weather Forecast., 32, 2175–2193, https://doi.org/10.1175/WAF-D-17-0038.1, 2017.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Mass, C. F., Baars, J., Wedam, G., Grimit, E., and Steed, R.: Removal of Systematic Model Bias on a Model Grid, Weather Forecast., 23, 438–459, https://doi.org/10.1175/2007WAF2006117.1, 2008.

Matthews, T., Perry, L. B., Koch, I., Aryal, D., Khadka, A., Shrestha, D., Abernathy, K., Elmore, A. C., Seimon, A., Tait, A., Elvin, S., Tuladhar, S., Baidya, S. K., Potocki, M., Birkel, S. D., Kang, S., Sherpa, T. C., Gajurel, A., and Mayewski, P. A.: Going to Extremes: Installing the World's Highest Weather Stations on Mount Everest, B. Am. Meteorol. Soc., 101, E1870–E1890, https://doi.org/10.1175/BAMS-D-19-0198.1y, 2020a.

Matthews, T., Perry, L. B., Lane, T. P., Elmore, A. C., Khadka, A., Aryal, D., Shrestha, D., Tuladhar, S., Baidya, S. K., Gajurel, A., Potocki, M., and Mayewski, P. A.: Into Thick(er) Air? Oxygen Availability at Humans' Physiological Frontier on Mount Everest, iScience, 23, 101718, https://doi.org/10.1016/j.isci.2020.101718, 2020b.

Matthews, T., Perry, B., Khadka, A., Sherpa, T. G., Shrestha, D., Aryal, D., Tuldahar, S., Thapa, N., Pradhananga, N., Athans, P., Sherpa, D. Y., Guy, H., Seimon, A., Elmore, A., Li, K., and Alexiev, N.: Weather Observations Reach the Summit of Mount Everest, B. Am. Meteorol. Soc., 103, E2827–E2835, https://doi.org/10.1175/BAMS-D-22-0120.1y, 2022.

McIlveen, J. F. R.: The everyday effects of wind drag on people, Weather, 57, 410–413, https://doi.org/10.1256/wea.29.02, 2002.

Minder, J. R., Mote, P. W., and Lundquist, J. D.: Surface temperature lapse rates over complex terrain: Lessons from the Cascade Mountains, J. Geophys. Res.-Atmos., 115, D14122, https://doi.org/10.1029/2009JD013493, 2010.

Miner, K. R., Mayewski, P. A., Baidya, S. K., Broad, K., Clifford, H., Elmore, A., Gajurel, A. P., Giri, B., Guilford, S., Hubbard, M., Jaskolski, C., Koldewey, H., Li, W., Matthews, T., Napper, I., Perry, L. B., Potocki, M., Priscu, J. C., Tait, A., Thompson, R., and Tuladhar, S.: An Overview of Physical Risks in the Mt. Everest Region, One Earth, 3, 547–550, https://doi.org/10.1016/j.oneear.2020.10.008, 2020.

Moore, G. and Semple, J.: Freezing and Frostbite on Mount Everest: New Insights into Wind Chill and Freezing Times at Extreme Altitude, High Alt. Med. Biol., 12, 271–275, https://doi.org/10.1089/ham.2011.0008, 2011.

Moore, G. W. K. and Semple, J. L.: Weather And Death On Mount Everest: An Analysis Of The Into Thin Air Storm, B. Am. Meteorol. Soc., 87, 465–480, https://doi.org/10.1175/BAMS-87-4-465y, 2006.

Napoli, A., Pepin, N., Palazzi, E., and Zardi, D.: A Workshop on Advances in Our Understanding of Elevation Dependent Climate Change, B. Am. Meteorol. Soc., 104, E928–E934, https://doi.org/10.1175/BAMS-D-23-0043.1y, 2023.

National Geographic: Mt Everest Weather Station Data, https://www.nationalgeographic.org/society/everest-weather-data/, last access: 29 October 2024.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting, J. Adv. Model. Earth Sy., 12, e2020MS002203, https://doi.org/10.1029/2020MS002203, 2020.

Thornton, J. M., Pepin, N., Shahgedanova, M., and Adler, C.: Coverage of In Situ Climatological Observations in the World's Mountains, Frontiers in Climate, 4, https://doi.org/10.3389/fclim.2022.814181, 2022.

Van Wyk de Vries, M.: AtsMOS: AtsMOS 1.0, Zenodo [code], https://doi.org/10.5281/zenodo.10889510, 2024.

Zhang, G., Zhu, S., Zhang, N., Zhang, G., and Xu, Y.: Downscaling Hourly Air Temperature of WRF Simulations Over Complex Topography: A Case Study of Chongli District in Hebei Province, China, J. Geophys. Res.-Atmos., 127, e2021JD035542, https://doi.org/10.1029/2021JD035542, 2022.