We thank both reviewers for their comments and their positive assessment of the manuscript. We have made a number of changes to our manuscript in response to the recommendations, in particular focusing on expanding and clarifying the methods section. We have added a new, more comprehensive methods flowchart, and additional methods table for the core machine learning modules, and additional details about different elements of the pre and post processing. We have also reviewed the full manuscript text and made small adjustments and corrections where recommended, which are highlighted in the 'track-changes' manuscript. We respond to all reviewer comments between the lines below.

Reviewer 2

This paper describes a model output statistics (MOS) approach aimed at improving weather forecasts in mountain environments. The authors focus on Mount Everest and show example usage with high elevation station data from the region. In addition to the model description, they explain their procedure for pre-processing the station data. I agree with the author's arguments on the importance of improved forecasts on Everest to improve climber safety and the need for combined NWP and in situ data to achieve this. The high altitude, telemetry equipped station network on Everest provides unique and interesting opportunities for model development targeting this goal. The example is instructive (though limited to one location) and the discussion section provides valuable context on strengths and limitations of the model. I agree with reviewer 1 regarding the need for a more detailed description of the model and I consider this the main area that should be revised/extended. I have a few additional suggestions below.

Thank you for your thorough review and constructive feedback. We appreciate your recognition of the importance of our work in improving weather forecasts for climber safety on Everest. We agree that a more detailed description of the model is necessary and will revise and extend this section accordingly. We will address the additional suggestions in detail below

General comments:

> Given the nature of the journal, I would like to see a more detailed description of the various model components. The model code is available but the documentation of the code could also be extended to help potential users get started. For the paper, I would suggest a subsection for each of the main processing modules, with particular focus on the steps in the "core processing" section of Fig. 1. This should include an explanation of the various learning techniques that were implemented.

We appreciate the feedback and have made substantial revisions to our manuscript to address these concerns. Recognizing the need for a more detailed description of the various model components, we have significantly expanded our methods section. This includes the addition of a new figure (detailing the workflow) and a table (outlining the specifics of the implemented ML models). The methods section has been reorganized into distinct subsections, each

focusing on the main processing modules, with particular emphasis on the steps in the "core processing" section of Fig. 1. We have provided detailed explanations of the learning techniques employed, such as XGBoost and Random Forest, including their parameter settings and implementation processes. Additionally, we have enhanced the documentation of the model code to assist potential users in getting started.

A subsection in the methods on the performance metrics used for model evaluation would also be beneficial.

We appreciate the suggestion and agree that a detailed subsection on the performance metrics used for model evaluation will enhance the clarity and comprehensiveness of our manuscript. We have added this subsection to the methods section ('2.4 Post-processing and validation metrics') to provide readers with a clear understanding of how we assess model performance.

Figures: Captions should be extended to fully explain the contents of the figures. Alternative visualizations for the contents of Figures 5-7 might be explored to better show the differences between learning methods.

We have modified figures 5-7 (now, figures 6-8 in the new manuscript) to better highlight the differences between the methods and capacity of the MOS technique to produce skillful forecasts. We show an example below. We have cropped the first line graph to only the region of overlap between data and model, provided an additional 'running error' plot to accompany this (b) and replaced the plot density graph with an error distribution histogram (d). Figure 8, showing the results for XGBoost, is shown below:
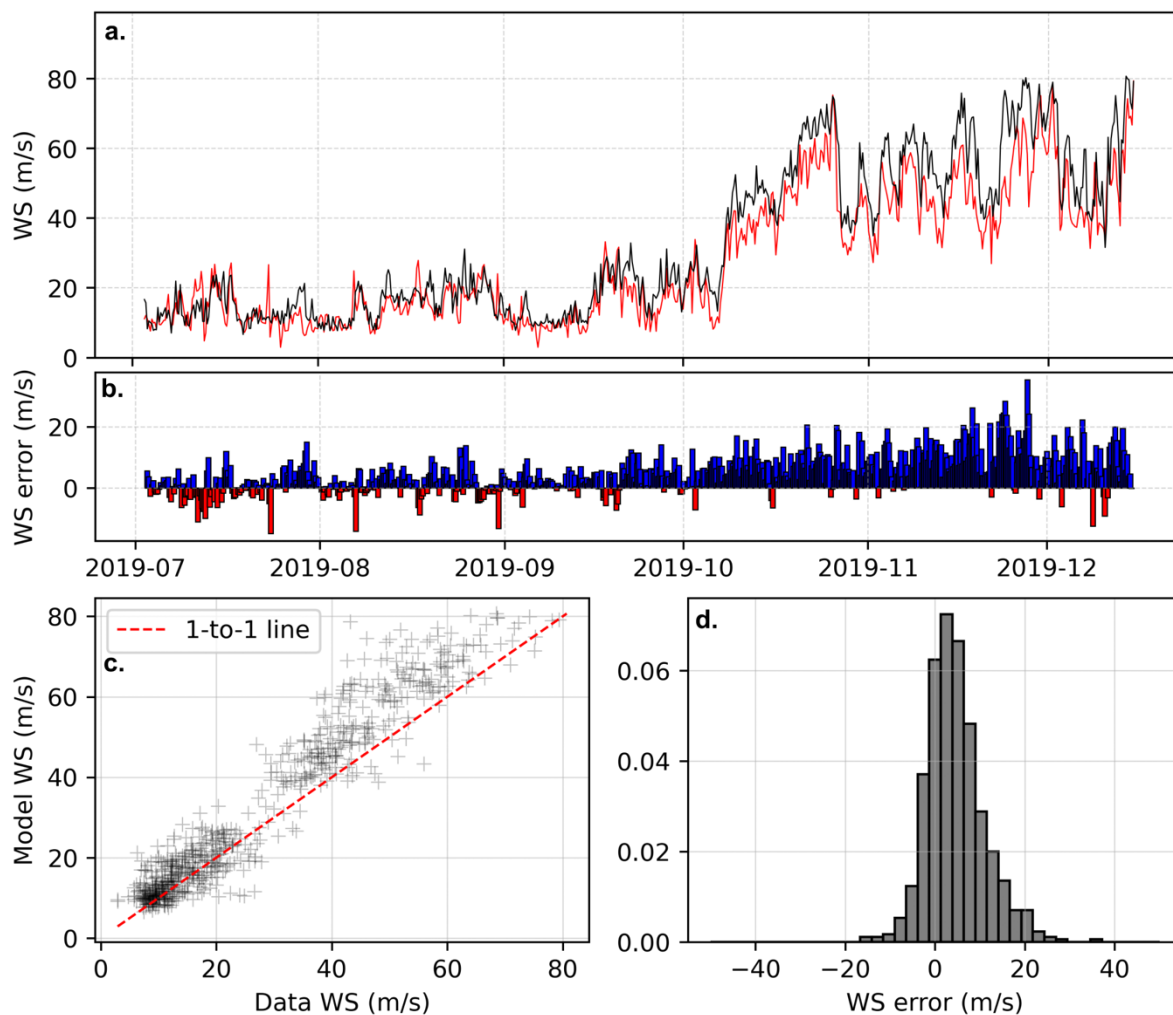
Specific comments:

Fig 4: What are the black and red lines? Add info in a legend, extend the figure caption to explain what is shown in the figure.

This caption has been extended and clarified to read "Validation of observational time series. Wind speed data shown in black (upper row) was considered reliable based on the dual-sensor correlation (lower row), while wind speed data show in red was judged unreliable and removed."

Fig 5-7: hard to see differences between the learning methods. Can this be combined to show the different model results in the same plot? What do the colors in the mesh plot (lower right panel) represent? Please explain in the caption, add a colorbar.

As discussed above, we have modified these three figures to better highlight these factors. We do not combine all three plots into 1 as this creates very small plots which are then challenging to interpret, but we have added two new plots which facilitate comparison: a running error plot and an error distribution histogram. Furthermore we provide a full range of

error evaluation metrics for this data, for instance for linear regression: "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."

Fig 8: Can the GFS forecast of temp and wind speed and the station data for the same time period be added for comparison? As is, the figure shows that the model outputs something and derivative values (WCT, frostbite time) can be computed. I believe more information could be added quite easily to enhance the contents of this figure.

We have added three new things to this figure:

-Comparative plots of the raw GFS forecasts for the Mt Everest summit (for wind speed and precipitation)

-The forecast probability of exceedance of our two wind speed thresholds (20 and 30 m/s).

-A classification of our forecast timeseries into 'hazard zones' based on these probability of wind speed threshold exceedances.

We hope that this clarifies this aspect of the manuscript.
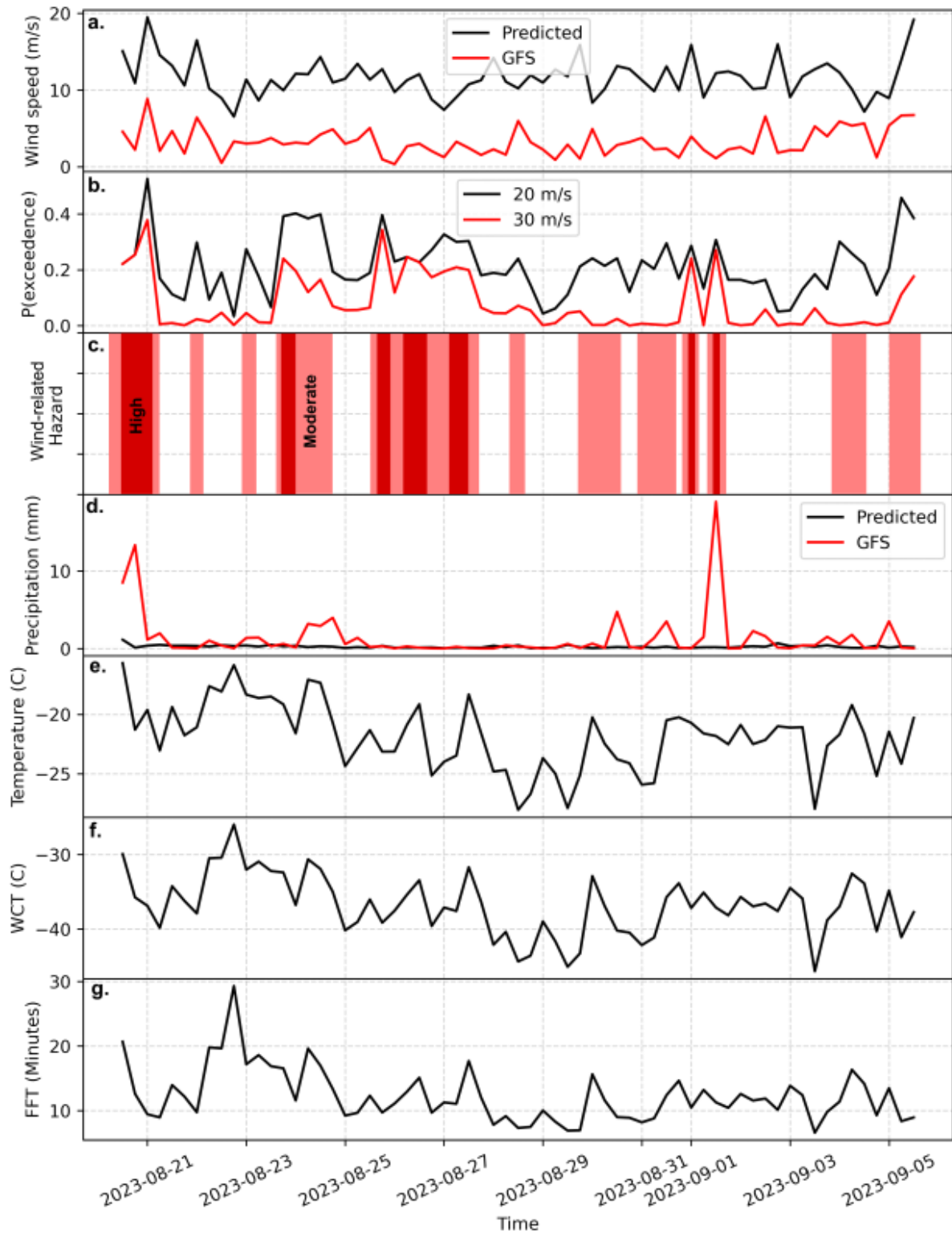
L184 This seems to be the first mention of Random Forest. In my opinion this should be introduced in the methods section.

We thank the reviewer for highlighting this and have added a new description of the Random Forest implementation to our methods section.

L187 Kling Gupta - This has also not been mentioned previously. Consider adding a subsection in the methods addressing your performance metrics.

As described in response to your query above, we have added a new 'performance metrics' paragraph to our methods section.

Typos:

L135 "Figure X"

This should refer to (now) Figure 5. We have corrected this.

L194 "the estimated are more closely clustered" - missing word?

This was a typo and should have read "the estimates are more closely clustered". We have corrected it.

L210 "The facial frostbit time briefly falls below 10 minutes this night also driven the the high wind speeds" – typos

We thank the reviewer for highlighting this issue and have corrected it to "The facial frostbite time briefly falls below 10 minutes this night also driven by the high wind speeds […]"