Reconstructing balloon-observed gravity wave momentum fluxes using machine learning and input from ERA5

Sothea Has[1], Riwal Plougonven[2], Aurélie Fischer[1], Raj Rani[3]
Francois Lott[3], Albert Hertzog[4], Aurélien Podglajen[3], Milena Corcos[5]

[1]CNRS/Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Université Paris Cité, France
[2]Laboratoire de Météorologie Dynamique, Ecole Normale Supérieure, IPSL, Paris, France
[3]Laboratoire de Météorologie Dynamique (LMD)/IPSL, PSL Research Institute, Paris, France
[4]LMD/IPSL, Sorbonne Université, Paris, France
[5]NorthWest Research Associates, Boulder, Colorado

Key Points:

- Eight superpressure balloons from the Strateole 2 mission provide observations for accurate gravity wave momentum flux estimation
- Three machine learning methods are employed to probe the relationship between the gravity wave momentum fluxes and ERA5's large-scale flow
- The most informative large-scale inputs are provided, along with a discussion of the successes and challenges of machine learning methods

Corresponding author: Sothea Has, `sothea.has@lpsm.paris`

Abstract

Global atmospheric models rely on parameterizations to capture the effects of gravity waves (GWs) on middle atmosphere circulation. As they propagate upwards from the troposphere, the momentum fluxes associated with these waves represent a crucial yet insufficiently constrained component. The present study employs three tree-based ensemble machine learning (ML) techniques to probe the relationship between large-scale flow and small-scale GWs within the tropical lower stratosphere. The measurements collected by eight superpressure balloons from the Strateole 2 campaign, comprising a cumulative observation period of 680 days, provide valuable estimates of the gravity wave momentum fluxes (GWMFs). Multiple explanatory variables, including total precipitation, wind, and temperature, were interpolated from the ERA5 reanalysis at each balloon's location. The ML methods are trained on data from seven balloons and subsequently utilized to estimate reference GWMFs of the remaining balloon. We observed that parts of the GW signal are successfully reconstructed, with correlations typically around 0.54 and exceeding 0.70 for certain balloons. The models show significantly different performances from one balloon to another, whereas they show rather comparable performances for any given balloon. In other words, limitations from training data are a stronger constraint than the choice of the ML method. The most informative inputs generally include precipitation and winds near the balloons' level. However, different models highlight different informative variables, making physical interpretation uncertain. This study also discusses potential limitations, including the intermittent nature of GWMFs and data scarcity, providing insights into the challenges and opportunities for advancing our understanding of these atmospheric phenomena.

Plain Language Summary

Part of the atmosphere's large-scale circulation results from motions that are not resolved, or partly resolved, by weather or climate models. These include internal gravity waves, with horizontal scales from a few to hundreds of kilometers. The main sources occur in the troposphere, such as flow over mountains and cloud development. Their three-dimensional propagation induces major aggregated impacts in the stratosphere and mesosphere, forcing key aspects of the circulation. This forcing is accounted for in climate models by 'parameterizations', that mimics the effect of the unresolved waves based on the large-scale, resolved flow. These parameterizations necessarily retain crude approximations and introduce significant uncertainty in the models. For GWs, sources are a major uncertainty. This study makes use of the high-altitude balloon campaign Strateole 2 (Oct. 2019-Feb. 2020). Eight balloons circled Earth at heights around 18 to 20 km, providing unique observations of the GWs. These are used as targets for machine learning (ML) methods that take as inputs the information from outputs of a numerical weather prediction model describing the large-scale flow. The successes and difficulties of ML provide insights which can guide improvements of parameterizations, such as the most informative large-scale variables for estimating the unresolved waves.

1 Introduction

Climate models and Numerical Weather Prediction models resolve a widening range of atmospheric processes as computing power increases, enabling finer spatial resolution. Subgrid-scale processes persist nonetheless, and efforts to improve and constrain them better are essential. Internal gravity waves constitute one of these subgrid-scale processes, with important implications for the circulation and variability of the middle atmosphere (Fritts & Alexander, 2003). Motivations for improved modeling of the stratosphere includes climate (e.g. Solomon et al. (2010); Kremser et al. (2016)) but also predictability on shorter time scales (F. Vitart and A.W. Robertson, 2018; Butchart, 2022).

Gravity waves occur on scales ranging from a few to several hundreds of kilometers. An important effect stems from their vertical propagation: gravity waves are responsible for vertical transfers of momentum from lower layers (troposphere: denser and with more gravity wave sources) to upper layers (stratosphere and beyond), where they constitute an essential driver of the overall circulation (Fritts & Alexander, 2003). A significant part of the spectrum of gravity waves has been and remains unresolved in global models, requiring these effects to be represented by parameterizations (Kim et al., 2003). Models display sensitivity to these, calling for coordinated efforts to better constrain these parameterizations from both observations and high-resolution modeling (Alexander et al., 2010).

A global comparison of observed, resolved and parameterized gravity wave momentum fluxes was carried out by Geller et al. (2013), highlighting significant discrepancies. Although GWs parameterizations are now used routinely in climate models, their validation against in situ observations remains a challenge. There exist global observations derived from satellite observations (e.g. Ern et al. (2018)), but there are limitations on the wavelengths that can be observed, and significant assumptions are needed to indirectly deduce important quantities like the momentum fluxes from temperature fluctuations, using polarization relations (Alexander et al., 2010; Ern et al., 2014). For these reasons superpressure balloons have been highlighted as a valuable and accurate source of information on gravity wave momentum fluxes (Geller et al., 2013). A downside of superpressure balloon observations is their very sparse sampling of the lower stratosphere: despite a broad coverage of the Southern Ocean (Jewtoukoff et al., 2015) and of the equatorial belt (Corcos et al., 2021), each balloon flight provides only local information: one time series along its trajectory.

There are fundamental difficulties in validating parameterizations of gravity waves: the purpose of a parameterization is to provide the forcing to the large-scale which is missing because of unresolved processes. Ideally, one would wish to *know* what this forcing should be and validate this outcome of parameterizations. Unfortunately, this forcing cannot be directly observed. Validating parameterizations by the realism of the climatology and variability of the atmospheric circulation in global models constitutes a first step, but is not a severe test and allows for compensating errors between parameterized processes (Plougonven et al., 2020). More stringent tests involve comparisons to observations (de la Camara et al., 2014; Trinh et al., 2016). Recently, direct comparisons between observed and parameterized gravity waves have been carried out on the scale of daily variations rather than at the level of general statistical characteristics (Lott et al., 2023). The large-scale environment was described using the ERA5 reanalyses (Hersbach et al., 2020), providing the background fields necessary to emulate the parameterization of convectively generated waves of Lott & Guez (2013), which is the parameterization used in the climate model of IPSL (Institut Pierre Simon Laplace, Boucher et al. (2020)). The comparison was quite encouraging, with the gravity wave momentum fluxes having the right order of magnitude, and an appropriate intermittency.

An essential aspect, and fundamental issue, to keep in mind when comparing observed and modeled gravity wave momentum fluxes is their strong intermittency: in time series of GWMF, one commonly finds short, intense peaks corresponding to a strong gravity wave event, surrounded by considerably weaker values. This has been highlighted in the long 'tail' of the Probability Density Function (PDF) of the GWMF (Alexander et al., 2010; Hertzog et al., 2012), and quantified in simulations and observations (Plougonven et al., 2013; Wright et al., 2013; Ern et al., 2022). This intermittency further contributes to making the parameterization of gravity waves a challenging task.

For the improvement of parameterizations in general (not only those of gravity waves), machine learning methods provide an array of possibilities. These have been explored in different directions:

- Machine learning can enable the emulation of parameterizations, leading to significant computational time savings (Chantry et al., 2021; de Burgh-Day & Leeuwenburg, 2023).
- Machine learning can help to capture the relationship between large-scale fields and the unresolved process, as illustrated in the case of convection by Gentine et al. (2018). For exploration, the dataset used as the truth came from a higher-resolution simulation, not from observations; obtaining observationally based knowledge of the effects to be parameterized remains a major challenge.
- Machine learning can be used to explore the relationship between the large-scale flow and the resulting small-scale waves, as has been done for orographic waves over Northern Japan (Matsuoka et al., 2020). Again, both the target and the inputs are modelled fields, but at different resolutions.
- As a precursor to a data-driven parameterization that would have learned from observations, a machine learning-based emulator of a parameterization for gravity waves has been used in a climate model, including under climate change conditions (Espinosa et al., 2022).

The purpose and scope of the present study is to probe the relationship between the large-scale flow and gravity waves in the Tropics, using machine learning approaches to address fundamental issues: what fraction of the GWMF can be determined from knowledge of the large-scale flow, and what fraction remains as *stochastic*? Which large-scale variables are most informative, and do they match with our common understanding of underlying gravity wave parameterizations? The present study belongs to the third category outlined above for the uses of machine learning (the purpose is *not* to produce a new parameterization, nor to emulate an existing one). With similar goals, Amiramjadi et al. (2023) used machine learning methods to probe the relationship between the large-scale flow and gravity waves, for non-orographic waves in the mid-latitudes and using waves resolved in a reanalysis as a target. In contrast, the present study aims at *observed* momentum fluxes in the Tropics, where the Strateole 2 campaigns provide a wealth of new observations (Haase et al., 2018; Corcos et al., 2021).

The paper is organized as follows: Section 2.1 provides an overview of the data and ML algorithms used in this study. Section 3 presents the performances of ML methods in reconstructing the reference GMWFs. Section 4 discusses the factors that influence the performances and addresses the limitations of ML methods. Finally, Section 5 concludes the study with key takeaways and future directions.

2 Data and methodology

2.1 Data

We use in situ observations collected from eight constant-level balloon flights (altitude between 18.5 and 20km) during the Strateole-2 mission from November 2019 to February 2020 (Corcos et al., 2021). As in Corcos et al. (2021), momentum fluxes (MFs) were computed from raw balloon measurements following the procedure described in Vincent and Hertzog (2014). Essentially, the pressure and horizontal wind time series are first projected in the time-frequency domain thanks to a continuous wavelet transform (Torrence and Compo, 1998). The pressure observations inform on the vertical displacements of the balloon, which are related to those of air parcels, assuming that the balloon behaves as a perfect isopycnic tracer. The time-frequency MF decomposition is then derived from the wavelet cross-spectrum of the horizontal winds and air-parcel vertical displacements. Segments polluted by non-geophysical artifacts (e.g. depressurization events) are discarded.

For our analysis, and following Corcos et al. (2021), we considered gravity wave MFs integrated over two frequency bands: a high-frequency (HF) band (i.e. short pe-

riods, ranging from 15 minutes to 1 hour) and wide-frequency (WF) band (i.e., long periods, ranging from 15 minutes to 1 day). For the sake of readability, in all that follows we focus on the HF band, unless explicitly stated. Additionally, we also differentiate between eastward-propagating waves that yield positive MF in the zonal direction (eastward) and westward-propagating waves that produce negative MF (westward). We use these MFs as a reference for the true target MFs. Then, we pair them with large-scale flow input information from ERA5, such as wind velocity (u and v), temperature (temp), total precipitation (tp) and logarithm of surface pressure (lnsp). These fields are retrieved for each balloon, from fields at a resolution of $1^\circ \times 1^\circ$, at the grid point closest to the balloon position. Additionally, the same input variables have been retrieved in the vicinity of a 5 by 5 horizontal square centered on the grid point closest to the balloon; in the present study, only total precipitation in this extended area around the balloon will be used. In the vertical, the ECMWF model comprises a total of 137 levels. Four levels are retained in the present study, to succinctly describe the vertical wind profile from the surface to balloon flight level (see Table 1).
The inputs and the targets are interpolated and averaged into 1-hour time resolution. The three ML models are trained using 3-hour time averaging data, and their performance will be evaluated based on daily averaging time resolution, as presented in Lott et al. (2023). Table 1 presents the finalized large-scale flow variables utilized for training ML models.

2.2 Methodology

In this study, three tree-based ensemble ML methods are considered: random forest (RF) introduced in Breiman (2001), extremely randomized trees also known as extra-trees (ET) by Geurts et al. (2006), and Adaptive Boosting or Adaboost regressors by Freund & Schapire (1997). These algorithms construct multiple decision trees, and the final prediction is determined by aggregating the individual decision tree predictions.

It should be noted that other methods, such as deep neural networks, as well as other types of networks including convolutional and recurrent neural networks, have also been implemented. However, the performances of these methods are not comparable to the presented tree-based algorithms, as these models typically require a large number of observations to achieve comparable results. The limitations and concerns regarding the models, the large-scale input variables, the target observations, and the nature of the relation between the large-scale and small-scale flow will be discussed later in Section 4.3.

2.2.1 Decision tree

The decision tree algorithm (Breiman et al., 1984) is the foundational building block of the primary ML methods used for our predictions. They are widely used for nonlinear prediction problems due to their efficiency and interpretability. To construct a decision tree, the training data is recursively partitioned into small hyperrectangular regions of the forms $R_1 = \{X \leq \alpha\}$ and $R_2 = \{X > \alpha\}$ for some ERA5 input variable X (wind velocity or precipitation, for instance) and threshold α. At each step, we recursively split the input space into hyperrectangular regions that are as pure as possible. Purity refers to the homogeneity of the training target y (GWMF) within each region, and Total Within Sum of Squares (TWSS) is utilized as the impurity measure in this study. Specifically, a split is performed at any input variable X at threshold α if it minimizes the following TWSS criterion:

$$\sum_{y \text{ of } R_1} (y - \mu_1)^2 + \sum_{y \text{ of } R_2} (y - \mu_2)^2,$$

where

- R_1 and R_2 are the left and right regions of the split

- μ_1 and μ_2 are the average targets within region R_1 and R_2 respectively.

Any new observation must belong to one of these regions, and its prediction is determined by averaging the target values of all the neighboring observations within that block. Constructing an optimal tree is generally challenging, and the tree's structure, such as its depth and the minimum size of regions allowed to split, are hyperparameters that need to be optimized. Figure 1 below provides an example of a simple decision tree trained on 100 observations of precipitation and zonal wind velocity to predict absolute GWMF.

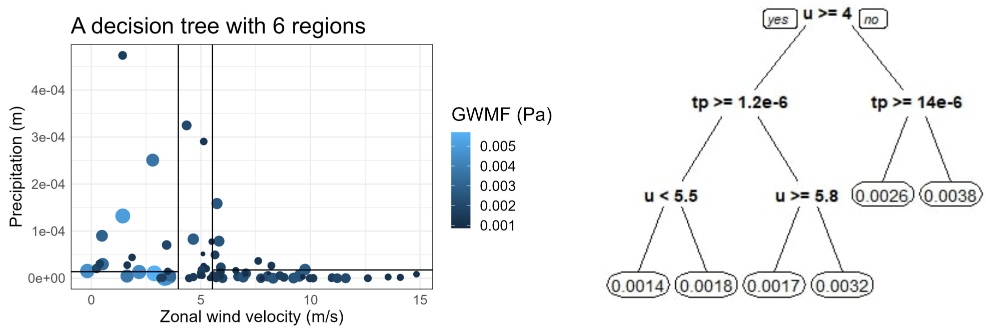

Figure 1. An example of a simple decision tree built using precipitation and wind velocity to predict absolute GWMF. The left side is the partition cell representation of the tree on the right side. The data points are colored according to the value of their target GWMF.

2.2.2 *Random forest*

Random forest (RF) is a powerful ensemble learning method that aims at minimizing variance across a collection of decision trees by averaging their predictions (Breiman, 2001). The term 'random' signifies the deliberate characteristic of constructing individual trees using different bootstrap samples (sampling observations with replacement) and exploring only a small, randomly selected, subset of the complete input features. This approach effectively decorrelates the individual trees, resulting in a reduction of prediction variance. Additionally, the construction of each individual tree using only a small subset of input features enables random forest to handle high-dimensional data effectively. The key hyperparameters in a random forest are the number of trees, tree complexity, and the number of randomly selected features used in building the individual trees. Fine-tuning these hyperparameters is essential to optimize the performance of the method.

2.2.3 *Extremely randomized trees*

Extremely randomized trees or Extra-trees (ET) operates similarly to RF approach, with the distinction that each tree is constructed using the complete training data, and each split is performed at *random values* using a *random subset* of input features (Geurts et al., 2006). This results in a high degree of independence among the trees and can occasionally yield remarkable results compared to the random forest method.

2.2.4 *Adaptive boosting*

Adaptive boosting (Adaboost) combines weak learners to create a strong predictive model (Freund & Schapire, 1997). Weak learners refer to predictive models that perform slightly better than random guesses, and simple decision trees with only a few splits (stumps) are used as weak learners in this study. During each iteration, Adaboost combines an individual stump by using a weighted sum, where the weight assigned to the current stump is determined based on its overall performance in predicting the target variable. Additionally, the weights associated with the individual training data points are adjusted manually based on their prediction accuracy, giving more attention or weight to points with poor predictions in the next iteration. Adaboost is well known for its ability to mitigate overfitting (Rätsch et al., 2001) and has achieved significant success in various prediction challenges (see, for example, Benjamin Bossan (2015) and ZEWEICHU (2019)).

2.2.5 K-fold cross validation

K-fold cross-validation is the most commonly used model selection technique in machine learning. It involves dividing the training data into K parts or folds, namely $F_1, \ldots, F_K$, then a model is trained on $K-1$ folds, and it is tested on the remaining one. This process is repeated K times and the final performance is the average performance over all the K different testing folds. In this study, K-fold cross-validation is used to prevent overfitting and to select the best possible hyperparameters of each ensemble method. More precisely, if f_θ is the considered method (random forest, for example) with a hyperparameter $\theta \in \Theta$, then the optimal hyperparameter θ^* is defined by,

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{K} \sum_{k=1}^{K} \sum_{(x_i, y_i) \in F_k} (f_\theta(x_i) - y_i)^2. \tag{1}$$

In our study, θ consists of the depth of the decision trees (maximum number of splits performed from the root node to the leaves), the size of random subsets of the ERA5 input features to be considered when building individual trees, and the number of decision trees used in each ensemble learning method. All these keys are tuned using 10-fold cross-validation.

2.3 Training

We first train ML models with an extensive set of ERA5 inputs. Subsequently, we refine these inputs to a more manageable subset (see Table 1 below) using importance feature scores, which will be described in Section 3. Moreover, in order to reduce the influence of extreme values in the target and increase its normality, the Box-Cox transformation (Box & Cox, 1964) is performed on the GWMF y to obtain the transformed target $\tilde{y}$:

$$\tilde{y} = \frac{y^\lambda - 1}{\lambda}.$$

In the experiment, the exponent $\lambda = 0.6$ is chosen based on the performance of models trained on the corresponding transformed target data. The predictions given by ML models are then reverted using the inverse transformation:

$$y = (1 + \lambda \tilde{y})^{1/\lambda}.$$

Moreover, to predict any GWMFs (absolute, eastward, or westward GWMFs of HF or WF case) of any given balloon, the ML models are trained using data from the seven other balloons. The models are fine-tuned using a 10-fold cross-validation method to optimize their performances.

Finally, the resolutions used for the data (see Section 2.1) reflect the phenomena we aim to estimate. From large-scale information as described from reanalyses at a resolution of $1^\circ \times 1^\circ$ and hourly in time, it is only reasonable to estimate GWMFs averaged over a comparable timescale (one hour). As the balloons drift at velocities typically ten to twenty m.s^{-1}, this corresponds to sampling over a spatial area of several tens of kilometers. The final choice for the specific setting used has been also guided by the motivation to make comparison with the results of Lott et al. (2023) possible.

The targeted gravity waves, as observed by the balloons, cover the whole range of intrinsic frequencies. The high frequency band (HF, see Section 2.1) may a priori be more difficult to predict from ML because it is expected to be more intermittent (Corcos et al., 2021), so that sampling will be a more severe issue than for the WF band. On the other hand, higher frequency waves propagate more vertically and are shorter-lived, both factors contributing to a stronger causal relationship between local conditions below the balloons and observed gravity at balloon level. As it has turned out that this second factor is more important, we focus hereafter on HF waves as the target, while the WF cases are detailed in the supplementary document.

Name	Notation	Description
Zonal, meridional wind velocity (m.s^{-1}) & temperature (K)	u_j, v_j & temp_j	with vertical level $j \in \{0, 2, 9, 19\}$ (km), where 0 is the surface and 19 is the balloon's level.
Total precipitation (m)	tp	at center of horizontal grid points.
Mean & standard deviation of precipitation (m)	$\mathrm{tp}_{\mathrm{mean}}$ & $\mathrm{tp}_{\mathrm{sd}}$	over horizontal grid points.
Solar zenith angle (°)	sza[1]	at the location of the balloon.
Log surface pressure (log(hPa))	lnsp	at the surface level.

Table 1. Large-scale input data for training ML models.

2.4 Evaluation metric

An important aspect in any comparison of models to observations is the choice of a metric to evaluate the performance of the models. We explain here why, in line with Lott et al. (2023), we use correlation between modelled and observed values as our metric. The current study is in line with studies that have compared parameterized and observed gravity waves (eg Geller et al. (2013)). In such comparisons, the first aim is naturally to compare *mean* momentum fluxes, yet over the past decade the importance of having a realistic variability has been emphasized (Alexander et al., 2010)). This has highlighted the notion of intermittency (Hertzog et al., 2012) and quantification of the distribution of momentum fluxes when comparing parameterizations to observations (de la Camara et al., 2014; Bushell et al., 2015). These comparisons, however, concern the overall statistics, not a direct comparison of observed and parameterized variations on a case-to-case basis. Obtaining an appropriate observational dataset and gathering the corresponding large-scale variables for such a case-to-case comparison has required significant work and has been achieved for the comparison of Lott et al. (2023). These datasets provide a unique opportunity to investigate the co-variability of observed GWMF and estimations from the large-scale flow, whether based on parameterizations (Lott et al., 2023)

[1] Solar zenith angle is the only input obtained from the balloons, not from the ERA5. It is a periodic function that provides an estimation of time of the day and the balloon's location.

or on machine learning techniques presented in this study. This is why we here focus on this co-variability, quantified by the correlation. It is expected that the averaging effect of tree-based algorithms may lead to underestimation of the target, especially when dealing with rare extreme values such as GWMFs. Obtaining appropriate intermittency of the reconstructed gravity wave momentum fluxes will require further efforts, and directions for these efforts are discussed in the perspectives (Section 5).

3 Results

This section reports the correlations of ML methods in reconstructing various types of observed GWMFs. The numerical study is carried out using `sklearn.ensemble` module in Python (Pedregosa et al., 2011). In general, the three ML models exhibit very comparable performances on any given balloon. In contrast, the performance of the ML models varies significantly from one balloon to another. At their best, ML models can achieve an encouraging level of correlation larger than 0.7. The average performance over all balloons and data exceeds 0.5. The worst performances is found for westward GWMF for a specific balloon, with correlation down to 0.2. Overall, the performances of ML models are sensitive to the choice of balloons and the types of GWs being considered (eastward, westward or absolute GWMFs). The numerical results for HF waves are presented in the following subsections, while the WF cases are presented in supplementary document.

3.1 Overall performances

Three examples of observed and predicted GWMFs of the HF case are presented in Figure 2 below. Each subplot displays the eastward component of the GWMFs in the positive part and the westward ones in the negative part. It can be observed that the models effectively capture the fluctuations of the observed momentum fluxes, particularly on balloon 2. However, the models struggle to fully estimate the amplitudes of high-peak events, especially for balloons 3 and 7. Overall, the performances of all ML models are quite similar; however, there are cases where one outperforms the others. For example, Adaboost appears to do a slightly better job on balloon 2 than the other two models in capturing the amplitudes of the high-peak events. It is worth noting that balloon 2 presents overall the best performance for the ML models, balloon 7 illustrates a typical average case, and balloon 3 is the most challenging one to predict: this is suggested visually in Figure 2, and is confirmed quantitatively in Table 2.

A feature of the reconstructed GWMF is that the peak values are generally underestimated, as can be seen even for balloon 2 in Figure 2. This is partly expected given that tree-based models involve averaging from numerous decision trees, some of which are insufficiently informed to capture extreme occurrences of GWMFs. To document the relationship between the reconstructed and observed GWMFs, scatterplots are displayed in Figure 3. These illustrate how the reconstruction captures well the variations of GWMFs, especially for rather weak variations. In contrast, for occurrences of larger MFs, the observed values cover a range of values that are not captured by the ML approaches. The scatterplots illustrate that those occurrences are rare, and the training data certainly constitutes a limiting factor. It is not clear that it may be possible to capture, in a deterministic way, these extremes. It is worth noting that the ML approaches do generally capture when the GWMF is at the high end of the range of reconstructed values.

Figure 4 presents boxplots of Pearson's correlation coefficients between predicted and true GWMFs of the HF case. Firstly, choosing the best model is challenging due to the variability in the boxplot positions, which depends on the choices of balloons and GWMF types. For instance, on balloon 2, the correlation boxplot of Adaboost is higher than the other two methods for the absolute and westward cases but lower than Random Forest for the eastward case. However, these differences are generally insignificant

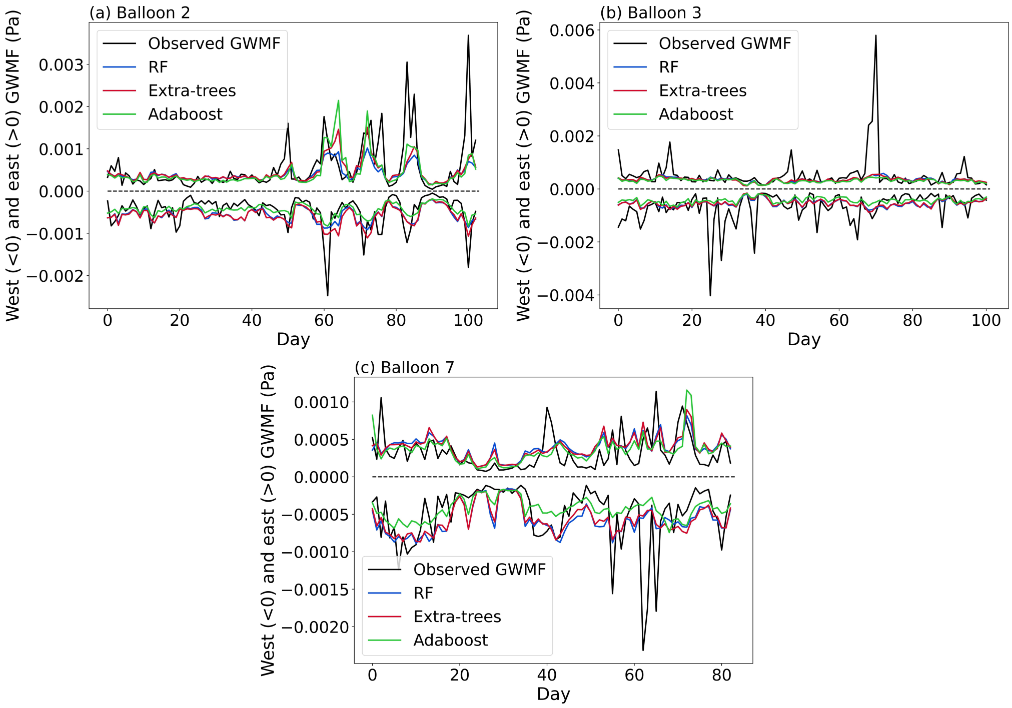

Figure 2. Observed and predicted time series of high-frequency east and westward GWMFs of the best, worst and medium cases: balloon 2, 3 and 7, respectively. The x-axis label "Day" indicates the number of days since the individual balloon was launched, with 0 corresponding to the moment of launch.

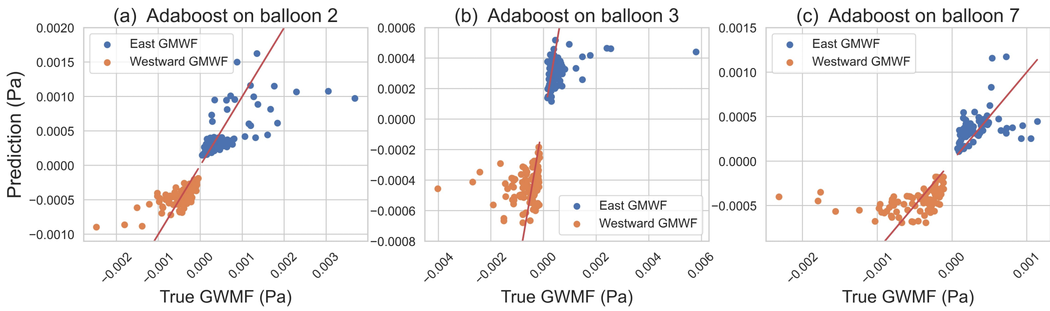

Figure 3. Scatterplots of predictions against observed (true) GWMF corresponding to the time series of Figure 2. Only the predictions of Adaboost are presented for balloon 2, 3 and 7 (from left to right). The lower groups represent the westward fluxes, while the upper groups denote the eastward ones. The red line serves as the reference 1:1 line.

compared to the variations observed between different balloons. Secondly, ML models demonstrate strong performance on balloons 2, 6, and 8 across all types of momentum fluxes, and they also excel in predicting the eastward momentum flux of balloon 1. Nevertheless, balloons 3, 4, 5, and 7 pose greater challenges, with the most difficult being the westward component of GWMF on balloon 3. Finally, the ML models generally outperform the gravity wave drag scheme of the IPSL model (Lott et al., 2023), except for balloon 3 (east and westward) and balloon 4. Moreover, Table 2 provides the statistical significance of the correlations presented in Figure 4.

Flight	Alt	Start	End	Duration/ DOF	Absolute			Eastward			Westward		
					RF	ET	AB	RF	ET	AB	RF	ET	AB
01_STR1	20.7	12/11/19	28/02/20	107/53	***0.56***	***0.57***	***0.58***	***0.67***	***0.69***	***0.67***	***0.38***	***0.37***	***0.43***
02_STR2	20.2	11/11/19	23/02/20	103/51	***0.70***	***0.67***	***0.74***	***0.67***	***0.62***	***0.65***	***0.60***	***0.63***	***0.70***
03_TTL3	19.0	18/11/19	28/02/20	101/33	***0.45***	***0.48***	***0.49***	**0.41**	***0.49***	***0.43***	0.21	0.23	0.18
04_TTL1	18.8	27/11/19	02/02/20	67/22	**0.44**	**0.43**	***0.47***	***0.47***	***0.48***	**0.44**	*0.35*	0.33	*0.37*
05_TTL2	18.9	05/12/19	23/02/20	79/19	**0.51**	***0.56***	***0.55***	*0.39*	**0.48**	0.35	0.35	*0.40*	**0.50**
06_STR1	20.5	06/12/19	01/02/20	57/10	***0.72***	***0.74***	***0.75***	**0.64**	**0.65**	**0.70**	**0.68**	***0.72***	*0.57*
07_STR2	20.2	06/12/19	28/02/20	83/16	**0.51**	**0.53**	**0.48**	*0.46*	**0.49**	*0.42*	*0.44*	*0.45*	0.32
08_STR2	20.2	07/12/19	22/02/20	77/12	***0.74***	***0.76***	***0.72***	***0.71***	***0.71***	***0.68***	***0.66***	***0.66***	**0.64**

Table 2. Average correlation coefficients between predicted and observed high-frequency GWMFs in 24h time resolution. In each case, by using decorrelated time as the degree of freedom (DOF), t-test statistics can provide the significance of each correlation with the convention: ***italic boldface*** = 99%, **boldface** = 95%, *italic* = 90%, and normal font = below 90% significant. For any given type of GWMF, the underlined correlations indicate the best performance of ML method on that target.

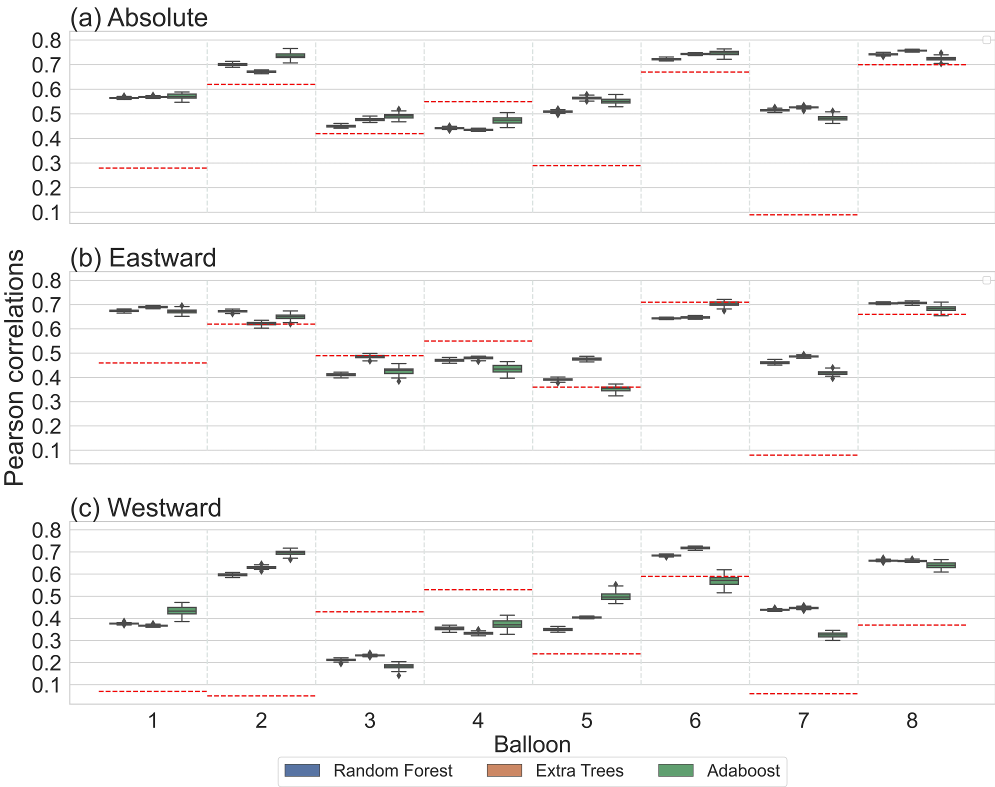

Figure 4. The boxplots display the correlations between predicted and observed high-frequency GWMFs obtained from 50 runs of ML methods as shown in Table 2. For each balloon, moving from left to right, the three boxplots correspond to the Random Forest, Extra Trees, and AdaBoost methods, respectively. The dashed horizontal red lines indicate the performance of the parameterization of the IPSL model (Lott et al., 2023).

3.2 Which large-scale inputs are informative for ML models?

The tree-based ensemble ML models employed in this study are not only proficient in predicting GWMFs but also offer valuable insights into the importance of large-scale input information during their training process. Each method exploits the feature importance (decrement of impurity measure at each split) of its individual decision trees for determining the overall feature importance, resulting in a ranking of input features from most to least important. Figure 5 showcases the ranking of the top 5 input features for all ML methods and GWMF types of the HF case.

Generally, the high-ranking inputs consist of variables that describe precipitation and wind velocity at and below the balloon's level. It is important to note that different models may not rank input features in the same way for a given target (as seen along the rows), due to the variations in the way individual trees are grown. However, the three models concur on the strongly impactful input features; for example, wind velocity at the balloon's level (u19) ranked first in the eastward case (second row) for all models. This suggests that the wind velocity surrounding the balloons is the most informative large-scale variable for predicting eastward gravity wave momentum fluxes (GWMFs). Furthermore, the few most significant inputs show a similar preference in both absolute and eastward GWMFs within the same model, as demonstrated in the columns of the first and second rows. For instance, standard deviation and average total precipitation

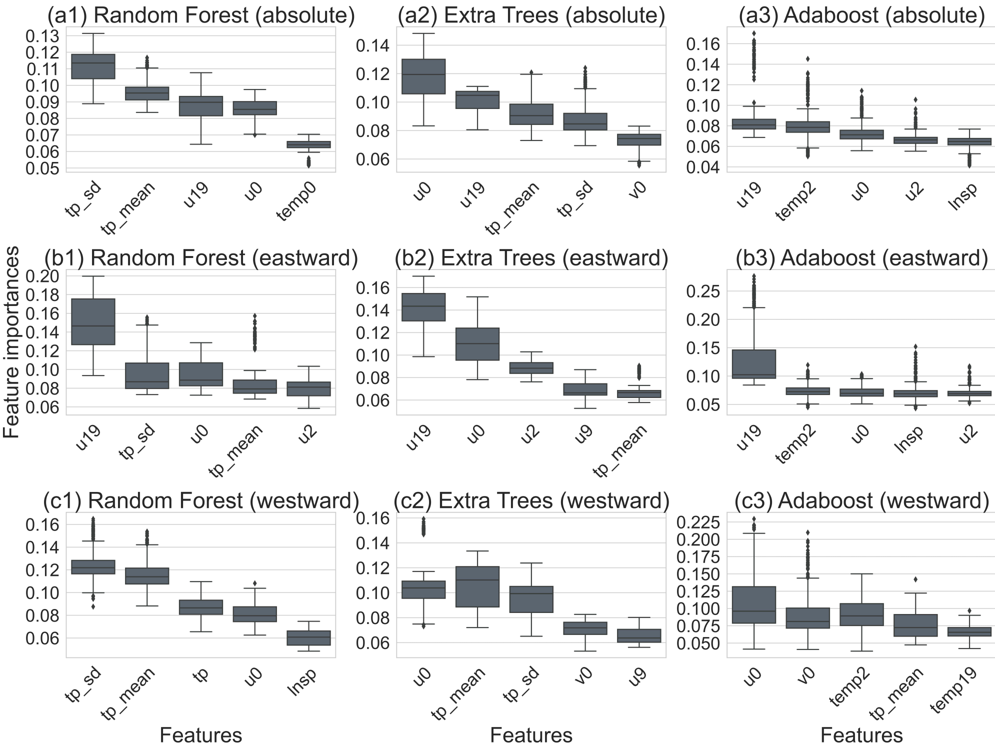

Figure 5. The boxplots show the 5 most important features given by different ML models (by column) on different types of targets (by row). Each boxplot is obtained from the same 50 simulations as displayed in Figure 4.

(tp_sd and tp_mean) are identified as impactful inputs in random forests, while surface zonal wind velocity (u0) is deemed the most important one in extra trees.

4 Discussion

While the results of the machine-learning models are generally encouraging, deficiencies and cases with poor performances were also found. The main motivation for this study being to probe the relationship between the large-scale and the unresolved process, these somewhat negative results are also of interest and can provide useful insights. Possible explanations for the main difficulties encountered are discussed below.

4.1 Why are westward GWMFs more challenging?

Figure 4 displays the performances of the ML models and those of the parameterization used in the IPSL climate model. Balloon 4 constitutes an exception, for which the parameterization systematically performs better than the ML methods. Leaving balloon 4 aside, ML approaches unambiguously outperform the parameterization for the absolute momentum fluxes. For the eastward momentum fluxes, ML approaches generally perform better or are similar to the parameterization. In contrast, both ML approaches and the parameterization have poorer performances for westward MF, and with greater variability for both: for five balloons, ML outperforms clearly the parameterization, whereas for two balloons (including balloon 4) the parameterization clearly outperforms the ML. The present section discusses possible reasons for this difficulty in reproducing the westward momentum fluxes.

Figure 6 displays the Probability Density Function of winds for three balloons as blue curves: balloon 2 has flown in winds that include a majority of westward, strong winds. Like balloon 1, it traveled near 10°S in easterly flow for a significant portion of its flight. In contrast, balloons 3 and 7 have flown in weaker winds, with a mild dominance of westerly winds. Also plotted in Figure 6 are conditional PDFs of the zonal winds, conditioned on the intensity of the absolute GWMF. The purpose is to detect if strong values of GWMF were associated to specific wind conditions. For balloon 2, strong GWMF values were found mostly for moderate to strong easterly winds, and this distribution is insensitive to the quantile chosen for the GWMF (90th, 95th or 99th percentile). For balloon 7, the distribution is somewhat sensitive to the quantile chosen. Finally, for balloon 3, the conditional distribution of zonal wind dramatically changes when it is restricted to the 99th percentile. This detects a particularly intermittent time series, with variability dominated by one extreme event, as seen from Figure 2. These findings contribute to explaining the poor performances for balloon 3: the variability of GWMF was dominated there by one (or very few) extreme events, occurring in a specific condition with very weak winds (close to zero, less than 5 m.s^{-1}). In contrast, the good performances for balloon 2 occur in a case with less intermittency, for which large GWMF are found in strong (easterly) winds.

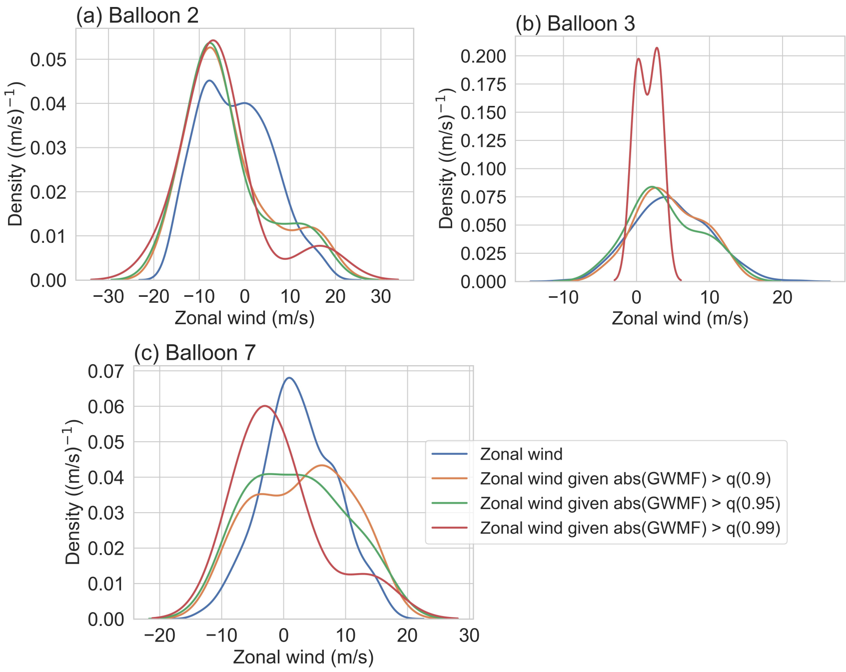

Figure 6. Conditional densities of zonal wind given different values of high-frequency westward GWMFs. Here, $q(0.9), q(0.95)$ and $q(0.99)$ are the 90%, 95% and 99% quantiles of the absolute value of high-frequency westward GWMFs, respectively.

From Table 2, Figure 4 and the trajectories of the balloons (Corcos et al., 2021), it appears that drifting with easterly winds may constitute a favorable factor (balloon 2), but neither a sufficient one (the correlation for westward momentum fluxes for balloon 1, which has a similar trajectory, is moderate, 0.43 at most) nor a necessary one: balloons 6 and 8 generally drift eastward, but good performances are found for the ML reconstruction of the westward MF (0.66 and 0.72 respectively).

Another aspect that influences the performances is the geographical location, and more specifically the latitude of the balloons. Figure 7 displays the PDF of latitude for the eight balloons, distinguishing those for which the ML reconstruction of westward MF is satisfactory (balloons 1, 2, 6 and 8, full lines) from those for which it remains challenging (balloons 3, 4, 5 and 7). Here again, one does not isolate a necessary condition, but the balloons for which reconstruction remain challenging are those that remain closest to the equator. This is consistent with the general expectation that dynamics is more complicated near the Equator, although it is not completely clear why this should matter for a small-scale process such as convectively generated gravity waves. It may be that it is not the dynamics itself that is intrinsically more difficult to capture at the Equator: it may be the input variables that are poorer, less accurate, very close to the Equator. It is known indeed that significant errors, in particular in the wind, are present in the reanalyses very near the Equator (Podglajen et al., 2014; Baker et al., 2014; Ern et al., 2023) and the errors are enhanced within a few degrees of the Equator (roughly between 8°S and 8°N).

4.2 Why are some balloons easier to predict than others?

Figure 7 indicates that the predictability of the observed GWMFs is influenced by the balloons' position, specifically, their distances from the equator. Balloons that traveled farther from the equator, primarily south (except for balloon 6, which also explored farther to the north), were found to be easier to predict. This tendency is observed for balloons 1, 2, 6, and 8 which are the well-predicted balloons. In contrast, the challenging balloons spent most of their time flying within a few degrees of the equator, where the atmospheric conditions are not well described by ERA5 data.

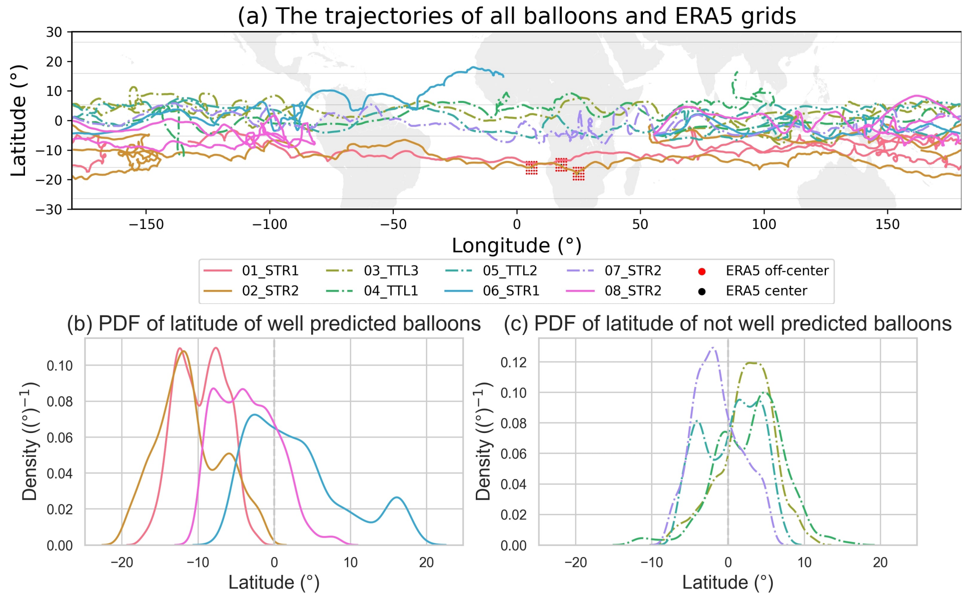

Figure 7. The trajectories of the balloons during the whole flight (a), and their latitude PDFs (b) and (c). Dashed lines correspond to balloons that pose challenges in prediction.

4.3 Exploring potential reasons for unsatisfactory cases

Several factors are expected to limit the ability to estimate the observed GWMFs from inputs describing the large-scale flow:

A. Part of the relationship between the large-scale flow and a subgrid-scale process such as gravity waves is non-deterministic, or stochastic: for given values of the large-scale fields, a range of different realizations of the subgrid-scale process is possible. It depends on the process: orographic gravity waves are likely more predictable than convective processes for instance.

B1. The estimate of GWMFs from superpressure balloons is very local and samples only along its trajectory. This is only partly mitigated by the hourly averaging. The GWMFs time series certainly remain sensitive to the specific location of the balloon. At present, it is difficult to estimate this sensitivity. Investigations with virtual balloons in high-resolution simulations shall be informative on this issue.

B2. A second concern regarding the target used for the ML is the observational error present in the estimates of the GWMFs from balloon measurements. These estimates are regarded as accurate because several variables are measured simultaneously and because of the quasi-Lagrangian nature of the measurements (Geller et al., 2013; Vincent & Hertzog, 2014). There remains nonetheless observational error.

C1. Concerns are also present for the input variables, and in particular it is known that the description of the equatorial dynamics is challenging, with significant errors remaining present in the reanalysis especially for wind (Podglajen et al., 2014).

C2. Another concern regarding input variables is that we may have omitted variables that could have been informative.

In our study, we mitigated the concern of omitting informative variables (C2.) by initially training ML models on a large set of ERA5 inputs, then selectively reducing them to a reasonably small subset, as described in Section 2.1. This approach ensures that essential ERA5 inputs are not inadvertently omitted. Furthermore, fine-tuning the hyperparameters of the models enhances their predictive capacity. Regarding the concern of large-scale variables (C1.), a sensitivity test to the error of ERA5's wind is described at the end of Section 5 (key messages).

In addition, we observe that all the balloons often flew over many convective processes, and the high-peak events often correspond to deep convective systems, as illustrated for selected cases in Figure 8 below. On January 12th, 2020, balloon 2 was flying in an area of convection (upper panels (a1) and (a2)), which is likely responsible for the highest peaks in its GWMF time series. Interestingly, for balloon 2, almost all events correspond very well with precipitation as described by ERA5 (first column of Figure 9). On the contrary, there is only one big event that happened for balloon 3 around January 29th, 2020 (lower left panel (b)). However, the ML models failed to capture it, as it appears to be absent from the ERA5 input variables (not reflected in precipitation nor winds as shown in the second column of Figure 9). This is also true for other challenging balloons, such as the 4th and 5th. Regarding balloon 7, the large-scale flow provide partial information for the high-peak events, resulting in partial success in the model's predictions.

5 Conclusion and perspectives

5.1 Key messages

The relationship between the large-scale atmospheric flow and gravity waves in the lower stratosphere has been investigated using Machine Learning (ML) approaches. This relationship is accounted for in global models through *parameterizations*. ML approaches allow us to revisit these in several ways, notably investigating how much of the subgrid-scale signal may be estimated *deterministically*, and which are the key variables for that purpose.

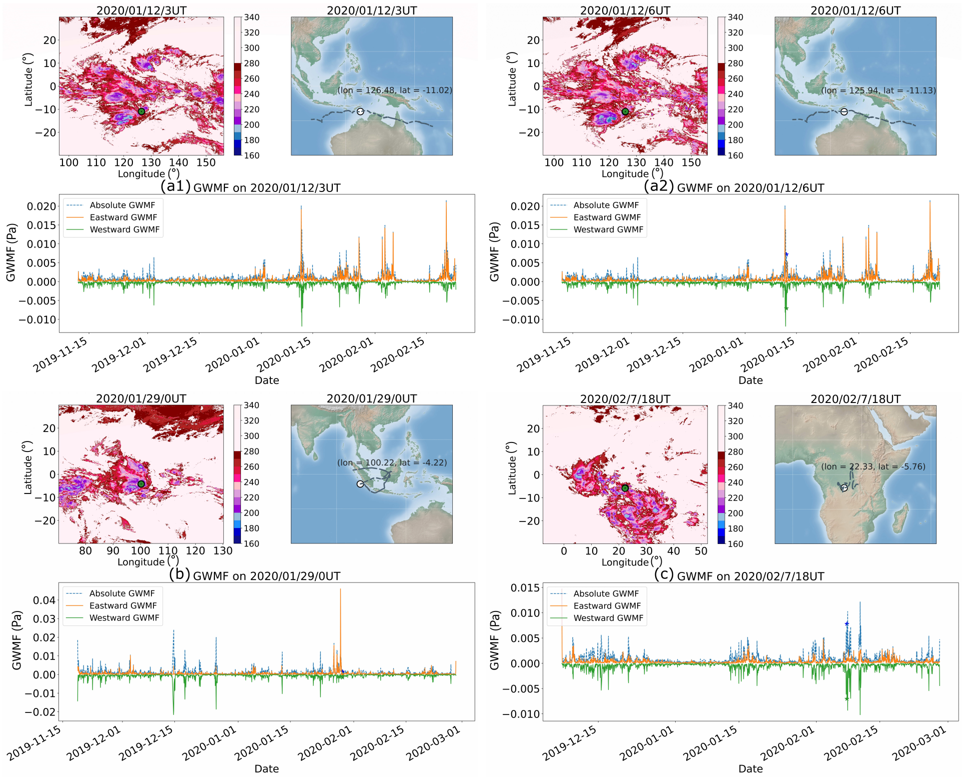

Figure 8. Brightness temperature from NOAA/NCEP GPM_MERGIR product (Janowiak, 2017), positions, and the corresponding observed GWMFs at the high-peak events of balloon 2 (top), balloon 3 (lower left) and balloon 7 (lower right).

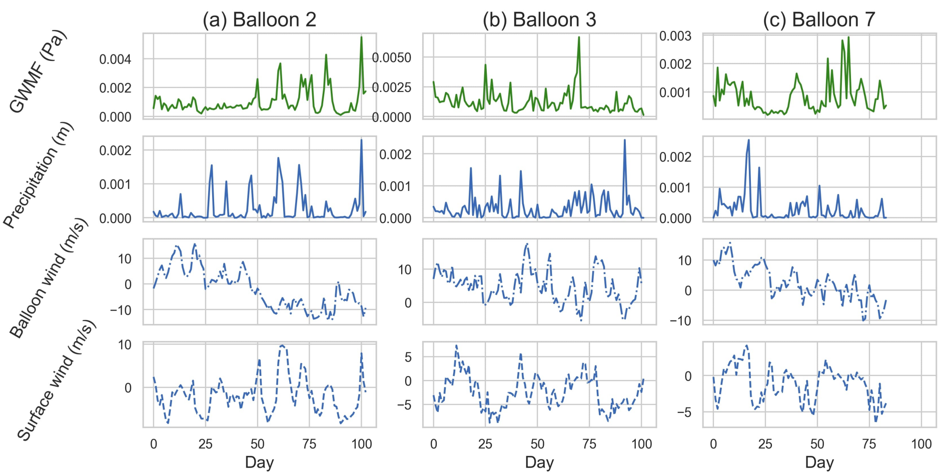

Figure 9. Time series of absolute GWMFs and the most informative ERA5 inputs in daily time resolution. The clear correspondence between precipitation and GWMF of balloon 2 can be visually observed in column (a). In contrast, this is not the case at all for balloon 3 as shown in column (b), and it partially presents in column (c) of balloon 7.

Estimates from superpressure balloon measurements were chosen as the target observations for gravity wave momentum fluxes (GWMF). The first campaign of the Strateole 2 project (Haase et al., 2018) consisted of eight balloons flying an average of about 85 days each around the globe in the equatorial band. The quasi-Lagrangian nature of the balloons allows an accurate estimate of gravity wave momentum fluxes (Geller et al., 2013), the latter being a key quantity for parameterizations (Alexander et al., 2010). Analysis of the GWMF estimated from measurements in this first campaign has highlighted and confirmed convection as the main source of gravity waves in this region, especially for waves with high frequencies (periods shorter than one hour); see Corcos et al. (2021).

The description of the large-scale flow environment was provided from the ERA5 reanalysis, along with vertical profiles co-located with each balloon at each time. These variables included wind, pressure, temperature, and precipitation. The latter being a noisy and uncertain field, values of total precipitation were retrieved in a 500×500 km^2 area around each balloon location, and was generally described by the mean and standard deviation over this area.

The ML models used are tree-based methods: random forests, extremely randomized trees, and adaptive boosting. Other methods were also investigated, as sensitivity experiments, without yielding major improvements. For each method, seven out of eight balloons were used for *training*, and the last balloon was used for *testing*.

The main results obtained from these investigations are as follows:

1. Based on the information provided by the large-scale flow data from ERA5, ML methods can reconstruct the observed GWMFs with correlations exceeding 0.7 in certain cases (balloon 2, 6, and 8), which is encouraging. Overall, the majority of the correlations are statistically significant at least at the 95% level, except for a few cases, as indicated in Table 2. The performances of ML methods, however, vary considerably from one balloon to another, with correlations down to 0.4 for some other balloons, and even down to 0.2 in one case. The overall average correlation for the HF case is 0.54, while a slightly lower average correlation of 0.49 is obtained in the WF case. In general, the correlations for WF waves are slightly weaker than those for HF waves (refer to the supplementary document for details).
2. The variations in performance are much larger between different balloons, than they are for a given balloon between ML approaches. This suggests that the performances are limited by the datasets, not by the choice of ML approach. The tree-based methods proved generally efficient, but there is not an overwhelming preference for one of them. Adaptive boosting frequently performed a bit better, but all three failed to capture the intensity of the (very intermittent) peaks in GWMF.
3. The most informative explanatory variables are those describing the precipitation and the zonal wind at and below the balloon's level. It is indeed an advantage of tree-based methods to provide information about the usefulness of the different inputs, e.g. through the Gini importance (Hastie et al., 2001). The importance of precipitation is consistent with the convective generation of the waves (Lott & Guez, 2013; Corcos et al., 2021). The importance of winds is consistent with the general understanding of the generation and propagation of waves (Kim et al., 2003); the relevance of wind at the balloon level is reminiscent of previous findings (Plougonven et al., 2017; Amiramjadi et al., 2023).
4. The ML methods were more efficient at reconstructing the part of GWMF associated with high-frequency waves (periods shorter than an hour) than the whole spectrum. This is consistent with the local character of the explanatory variables provided as inputs: high-frequency waves will be shorter-lived and propagate more vertically.
5. Different decompositions of the GWMF were used: absolute, eastward and westward GWMF. Interestingly, the performances significantly differed between these. The most difficult to reconstruct was found to be westward GWMF. Reasons for this likely include limitations of the dataset, to be further discussed below.

However, there are still parts where the large-scale flow are not informative enough in the estimation. There are cases where high peaks are present in the observed target, which indicates interesting events; however, large-scale flow are missed to describe them. As a result, the models failed to reconstruct such events in GWMFs (balloon 1 and 3, for example).

In addition, we have also implemented ML models by replacing ERA5's winds with balloon-observed winds at the balloon's level. This tests the sensitivity to errors in the input variables, for the variables for which we have direct observations, and which is known in the reanalysis to include significant error. The results suggest there is some sensitivity, but it is not extensive. Overall, the performances on some challenging balloons such as balloon 3 and 5 are significantly improved when using observed winds instead of ERA5's winds. In contrast, the performance on balloon 8 drops quite a bit compared to the model with ERA5's winds. Overall, the models utilizing observed wind achieve an average correlation of 0.53 in the HF case and 0.47 in the WF case. These results can be found in the supplementary document.

5.2 Perspectives

Although the ML approaches have performed well, and nearly always better than the parameterization, there are clear limitations to the current investigation, calling for further research. The very strong sensitivity of the performances to the balloon that is left out and then used for testing is a clear indication that we lack data: the results strongly depend on the split of the data for training and testing, the performances are far from convergence. This is consistent with the strong intermittency of the GWMF (Hertzog et al., 2012; Plougonven et al., 2013) and with the illustrative time series of Figure 2: for each balloon, GWMF are dominated by a few events, such that even with 680 days of balloon measurements, only a few handfuls of GWMF peaks are described. This is too little for data-driven methods. This also explains why clear distinctions between the different methods are not found: the ML methods do their best but still lack data to clearly separate a better method for this problem, if there is one.

Ways forward include:

- Obtaining more observations to use as the target, keeping the same framework for the ML. Additional observations would come from the second Strateole 2 campaign (in 2021) and from Loon balloons (Schoeberl et al., 2017; Köhler et al., 2023). The additional Strateole data would enhance the data by less than a factor 2 and is therefore not expected to suffice to make a dramatic change. The Loon data would come with other difficulties as the observations were not made for research purposes and come with their own challenges (Green et al., 2023).
- Additional data could be provided not for the targets, but for the explanatory variables. A first step could be including additional input variables from the reanalyses. However, preliminary attempts have not suggested significant gains from the most evident additional culprits. A second step would consist of providing much more detailed and more accurate information about the background flow: this could be obtained from satellite observations, such as the observations of brightness temperatures from geostationary satellites shown in Figure 8. This would constitute a very interesting new study but in a profoundly new framework and with different aims: to fully use the information available from satellites would a priori require providing maps (or images, or 2D fields) as input variables (more akin to Matsuoka et al. (2020), although their inputs were from models, not observations). The ML used would need to be reassessed (Matsuoka et al. (2020) used neural networks, for instance). Such a study would be of great interest because the performance of the ML methods would much less be tainted by the uncertainty (or errors) present in the inputs that serve to describe the background. Additionally,

much more detailed information would be provided about the background flow, allowing the ML methods to tap into a greater reservoir of potentially relevant information, and hence providing more precise answers regarding the relationship of the large-scale flow to the gravity wave signal. However, if the outcome of such an exercise would be of interest fundamentally, it would be more removed from the framework in which current parameterizations operate.

- A shortcoming of the present ML approaches is that they underestimate the peak values for GWMF (see Figures 2 and 3. This is expected, given the averaging involved in tree-based method and the limited number of strong events present in the training data. However, this implies that the distribution of reconstructed momentum fluxes misses the tail of intense, rare events, which are known to matter for atmospheric gravity waves (Hertzog et al., 2012; de la Camara et al., 2016). One way to overcome this would be to aim not at a deterministic reconstruction of the momentum fluxes, but at reconstructing a probability density function of these. This change of framework, equivalent to changing from a deterministic to a stochastic parameterization, would in fact be more consistent for three reasons: first, given some large-scale conditions, there are certainly several different small-scale configurations with different resulting gravity waves that can occur. Second, for any given realization of the small-scale flow corresponding to large-scale conditions, our observed values depend on the specific sampling by the balloon. At present, we do not fully know how sensitive the observed gravity wave momentum fluxes are to this sampling. Finally, the estimate of gravity wave momentum fluxes from the observed balloon measurements involves assumptions and methodological choices, and there is as always an observational error in the estimates for GWMF. Given that the ML methods do capture rather well the occurrence of larger values, using ML methods to reconstruct a PDF of likely fluxes, rather than a single, deterministic value, could give room to better represent the observed GWMF, although only in a probabilistic way.
- A fourth way forward consists in applying similar investigations on datasets where more data is available, albeit at the cost of more uncertainty on the realism of the data. High-resolution models such as global convection permitting simulations (Stephan et al., 2019) provide a wealth of information on the resolved gravity wave field, and many studies have repeatedly highlighted the ability of models to simulate efficiently many features of the observed gravity wave field (Plougonven & Teitelbaum, 2003; Wu & Eckermann, 2008; Preusse et al., 2014; Stephan et al., 2019). Model output from global simulations would provide amounts of data for which the sampling limitations of the Strateole balloons would not be present. The downside is the limitations of model data, relative to observations, and the need for strategies to validate which aspects of the simulations are realistic.

Acknowledgments

This work and Sothea Has are supported by the Institut des Mathématiques pour la Planète Terre (IMPT). This work has also received support from the ANR project BOOST3R (ANR-17-CE01-0016-01) and the French-American project Strateole 2 (CNES). Moreover, we gratefully acknowledge the support and collaborative efforts extended by members of the DataWave consortium, a Virtual Earth System Research Institute (VESRI) Schmidt Futures project.

6 Open research

Balloon data used in this study are presented in Haase et al. (2018) of the STRATEOLE 2 mission and can be extracted from the following website: `https://webstr2.ipsl.polytechnique.fr`. The ERA5 input variables are described in Hersbach et al. (2020) and can be obtained from the COPERNICUS access hub using the following website: `https://scihub.copernicus.eu/`. The machine learning algorithms implemented in our analysis are available in the `scikit-learn` Python library (Pedregosa et al., 2011) and can be downloaded from its website: `https://scikit-learn.org/stable/install.html`. Finally, the source codes for implementing machine learning methods in our analysis are made available at the following GitHub repository: `https://github.com/hassothea/Reconstruction\_of\_GWMF\_using\_ML\_ERA5`.

References

Alexander, M., Geller, M., McLandress, C., Polavarapu, S., Preusse, P., Sassi, F., ... Watanabe, S. (2010, July). Recent developments in gravity-wave effects in climate models and the global distribution of gravity-wave momentum flux from observations and models. , *136*, 1103-1124.

Amiramjadi, M., Plougonven, R., Mohebalhojeh, A. R., & Mirzaei, M. (2023). Using machine learning to estimate nonorographic gravity wave characteristics at source levels. *Journal of the Atmospheric Sciences*, *80*(2), 419–440.

Baker, W. E., Atlas, R., Cardinali, C., Clement, A., Emmitt, G. D., Gentry, B. M., ... others (2014). Lidar-Measured Wind Profiles: The Missing Link in the Global Observing System. *Bull. Am. Meteor. Soc.*, *95*(10.1175/2010JAS3455.1), 543-564.

Benjamin Bossan, W. K., Josef Feigl. (2015). *Otto group product classification challenge.* Kaggle. Retrieved from `https://kaggle.com/competitions/otto-group-product-classification-challenge`

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., ... others (2020). Presentation and evaluation of the ipsl-cm6a-lr climate model. *Journal of Advances in Modeling Earth Systems*, *12*(7).

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252. doi: 10.1111/j.2517-6161.1964.tb00553.x

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

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. *Wadsworth International Group*.

Bushell, A. C., Butchart, N., Derbyshire, S. H., Jackson, D. R., Shutts, G. J., Vosper, S. B., & Webster, S. (2015). Parameterized gravity wave momentum fluxes from sources related to convection and large-scale precipitation processes in a global atmosphere model. *Journal of the Atmospheric Sciences*, *72*(11), 4349–4371.

Butchart, N. (2022). The stratosphere: a review of the dynamics and variability. *Weather and Climate Dynamics*, *3*(4), 1237–1272.

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I., & Palmer, T. (2021). Machine learning emulation of gravity wave drag in numerical weather forecasting. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002477.

Corcos, M., Hertzog, A., Plougonven, R., & Podglajen, A. (2021). Observation of gravity waves at the tropical tropopause using superpressure balloons. *Journal of Geophysical Research: Atmospheres*, *126*(15), e2021JD035165.

de Burgh-Day, C. O., & Leeuwenburg, T. (2023). Machine learning for numerical weather and climate modelling: a review. *Geoscientific Model Development*, *16*(22), 6433–6477. Retrieved from `https://gmd.copernicus.org/articles/16/6433/2023/` doi: 10.5194/gmd-16-6433-2023

de la Camara, A., Lott, F., & Hertzog, A. (2014). Intermittency in a stochastic parameterization of nonorographic gravity waves. *J. Geophys. Res. Atmos.*, *119*, 11,905-11,919. doi: 10.1002/2014JD022002

de la Camara, A., Lott, F., Jewtoukoff, V., Plougonven, R., & Hertzog, A. (2016). On the gravity wave forcing during the southern stratospheric final warming in LMDz. *J. Atmos. Sci.*, *73*, 3213-3226. doi: 10.1175/JAS-D-15-0377.1

Ern, M., Diallo, M. A., Khordakova, D., Krisch, I., Preusse, P., Reitebuch, O., ... Riese, M. (2023). The quasi-biennial oscillation (qbo) and global-scale tropical waves in aeolus wind observations, radiosonde data, and reanalyses. *Atmospheric Chemistry and Physics*, *23*(16), 9549–9583. Retrieved from `https://acp.copernicus.org/articles/23/9549/2023/` doi: 10.5194/acp-23-9549-2023

Ern, M., Ploeger, F., Preusse, P., Gille, J., Gray, L. J., Kalisch, S., ... Riese, M. (2014). Interaction of gravity waves with the QBO: A satellite perspective. *Journal of Geophysical Research: Atmospheres*, *119*, 2329 - 2355. doi: 10.1002/2013JD020731

Ern, M., Preusse, P., & Riese, M. (2022). Intermittency of gravity wave potential energies and absolute momentum fluxes derived from infrared limb sounding satellite observations. *Atmospheric Chemistry and Physics*, *22*(22), 15093–15133. Retrieved from `https://acp.copernicus.org/articles/22/15093/2022/` doi: 10.5194/acp-22-15093-2022

Ern, M., Trinh, Q. T., Gille, P. P. J., Mlynczak, M., Russell, J., & Riese, M. (2018). GRACILE: a comprehensive climatology of atmospheric gravity wave parameters based on satellite limb soundings . *Earth System Science Data*, *10*, 857-892. doi: 10.5194/essd-10-857-2018

Espinosa, Z. I., Sheshadri, A., Cain, G. R., Gerber, E. P., & DallaSanta, K. J. (2022). Machine learning gravity wave parameterization generalizes to capture the qbo and response to increased co2. *Geophysical Research Letters*, *49*(8), e2022GL098174.

F. Vitart and A.W. Robertson (Ed.). (2018). *Sub-seasonal to seasonal prediction.* Elsevier.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Fritts, D., & Alexander, M. (2003). Gravity wave dynamics and effects in the middle atmosphere. *Reviews of Geophysics*, *41*(1), 1003. doi: doi:10.1029/2001RG000106

Geller, M., Alexander, M., Love, P., Bacmeister, J., Ern, M., Hertzog, A., ... Zhou, T. (2013). A comparison between gravity wave momentum fluxes in observations and climate models. *J. Clim.*, *26*, 6383-6405. doi: 10.1175/JCLI-D-12-00545.1

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophys. Res. Lett.*, *45*, 5742-5751. doi: 10.1029/2018GL078202

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42.

Green, B., Sheshadri, A., Alexander, M. J., Bramberger, M., & Lott, F. (2023). Gravity wave momentum fluxes estimated from project loon balloon data. *Authorea Preprints*.

Haase, J., Alexander, M., Hertzog, A., Kalnajs, L., Deshler, T., Davis, S., ... Venel, S. (2018, March). Around the world in 84 days. *EOS*, *99*. doi: 10.1029/2018EO091907

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning.* New York, NY, USA: Springer New York Inc.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049. Retrieved from `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803` doi: https://doi.org/10.1002/qj.3803

Hertzog, A., Alexander, M., & Plougonven, R. (2012). On the probability density functions of gravity waves momentum flux in the stratosphere. *J. Atmosph. Sci.*, *69*, 3433-3448.

Janowiak, J. B. . X. P., J. (2017). *Ncep/cpc l3 half hourly 4km global (60s - 60n) merged ir v1.* In A. Savtchenko, & M. D. Greenbelt (Eds.), Goddard Earth sciences data and information services center (GES DISC). (Accessed 2020).

Jewtoukoff, V., Hertzog, A., Plougonven, R., de la Camara, A., & Lott, F. (2015). Gravity waves in the Southern Hemisphere derived from balloon observations and ECMWF analyses. *J. Atmos. Sci.*, *72*, 3449-3468.

Kim, Y.-J., Eckermann, S., & Chun, H.-Y. (2003). An overview of the past, present and future of gravity-wave drag parametrization for numerical climate and weather prediction models. *Atmosphere-Ocean*, *41*, 65-98.

Köhler, L., Green, B., & Stephan, C. C. (2023). Comparing loon superpressure balloon observations of gravity waves in the tropics with global storm-resolving models. *Journal of Geophysical Research: Atmospheres*, *128*(15), e2023JD038549.

Kremser, S., Thomason, L. W., von Hobe, M., Hermann, M., Deshler, T., Timmreck, C., … others (2016). Stratospheric aerosol—observations, processes, and impact on climate. *Reviews of Geophysics*, *54*(2), 278–335.

Lott, F., & Guez, L. (2013). A stochastic parameterization of the gravity waves due to convection and its impact on the equatorial stratosphere. , *118*, 8897-8909.

Lott, F., Rani, R., Podglajen, A., Codron, F., Guez, L., Hertzog, A., & Plougonven, R. (2023). Direct comparison between a non-orographic gravity wave drag scheme and constant level balloons. *Journal of Geophysical Research: Atmospheres*, *128*(4), e2022JD037585.

Matsuoka, D., Watanabe, S., Sato, K., Kawazoe, S., Yu, W., & Easterbrook, S. (2020). Application of deep learning to estimate atmospheric gravity wave parameters in reanalysis data sets. *Geophysical Research Letters*, *47*(19), e2020GL089436.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Plougonven, R., de la Camara, A., Hertzog, A., & Lott, F. (2020). How does knowledge of atmospheric gravity waves guide their parameterizations? *Q.J. Roy.Meteorol. Soc.*, 1-15. doi: 10.1002/qj.3732

Plougonven, R., Hertzog, A., & Guez, L. (2013, January). Gravity waves over Antarctica and the Southern Ocean: consistent momentum fluxes in mesoscale simulations and stratospheric balloon observations. *Quart. J. Roy. Meteorolog. Soc.*, *139*, 101-118.

Plougonven, R., Jewtoukoff, V., de la Camara, A., Hertzog, A., & Lott, F. (2017). On the relation between gravity waves and wind speed in the lower stratosphere over the Southern Ocean. *J. Atmos. Sci.*, *74*, 1075-1093. doi: 10.1175/JAS-D-16-0096.1

Plougonven, R., & Teitelbaum, H. (2003). Comparison of a large-scale inertia-gravity wave as seen in the ECMWF and from radiosondes. *Geophys. Res. Let.*, *30*(18), 1954.

Podglajen, A., Hertzog, A., Plougonven, R., & Zagar, N. (2014). Assessment of the accuracy of (re)analyses in the equatorial lower stratosphere. , *119*, 11,166-11,188. doi: 10.1002/2014JD021849

Preusse, P., Ern, M., Bechtold, P., Eckermann, S., Kalisch, S., Trinh, Q., & Riese, M. (2014). Characteristics of gravity waves resolved by ECMWF. *Atmos. Chem. Phys.*, *14*, 10483-10508. doi: 10.5194/acp-14-10483-2014

Rätsch, G., Onoda, T., & Müller, K.-R. (2001). Soft margins for adaboost. *Machine Learning*, *42*.

Schoeberl, M. R., Jensen, E., Podglajen, A., Coy, L., Lodha, C., Candido, S., & Carver, R. (2017). Gravity wave spectra in the lower stratosphere diagnosed from project loon balloon trajectories. *Journal of Geophysical Research: Atmospheres*, *122*(16), 8517–8524.

Solomon, S., Rosenlof, K., Portmann, R., Daniel, J., Davis, S., Sanford, T., & Plattner, G.-K. (2010, March). Contributions of stratospheric water vapor to decadal changes in the rate of global warming. *Science*, 1219-1223. doi: 10.1126/science.118248

Stephan, C., Strube, C., Klocke, D., Ern, M., Hoffmann, L., Preusse, P., & Schmidt, H. (2019). Gravity Waves in Global High-Resolution Simulations With Explicit and Parameterized Convection. *J. Geophys. Res.*, *124*, 4446-4459. doi: 10.1029/2018JD030073

Trinh, Q., Kalisch, S., Preusse, P., Ern, M., Chun, H., Eckermann, S., ... Riese, M. (2016). Tuning of a gravity wave source scheme based on HIRDLS observations. *Atmos. Chem.Phys.*, *16*, 7335-7356. doi: 10.5194/acp-16-7335-2016

Vincent, R., & Hertzog, A. (2014). The response of superpressure balloons to gravity wave motions. *Atmospheric Measurement Techniques*, *7*(4), 1043–1055.

Wright, C., Osprey, S., & Gille, J. (2013). Global observations of gravity wave intermittency and its impact on the observed momentum flux morphology. *J. Geophys. Res.*, *118*, 10,980-10,993. doi: 10.1002/jgrd.50869

Wu, D., & Eckermann, S. (2008). Global gravity wave variances from Aura MLS: characteristics and interpretation. , *65*(12), 3695-3718.

ZEWEICHU. (2019). *2019 ttic 31020 hw4 spam (adaboost).* Kaggle. Retrieved from `https://kaggle.com/competitions/2019-ttic-31020-hw4-spam-adaboost`