Aviation Turbulence Forecasting at Upper Levels with Machine Learning Techniques Based on Regression Trees

DOMINGO MUÑOZ-ESPARZA, ROBERT D. SHARMAN, AND WIEBKE DEIERLING

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 16 May 2020, in final form 21 August 2020)

ABSTRACT: We explore the use of machine learning (ML) techniques, namely, regression trees (RT), for the purpose of aviation turbulence forecasting at upper levels [20–45 kft ($\sim 6\text{--}14$ km) in altitude]. In particular, we develop a series of RT-based algorithms that include random forests (RF) and gradient-boosted regression trees (GBRT) methods. Numerical weather prediction model prognostic variables and derived turbulence diagnostics based on 6-h forecasts from the 3-km High-Resolution Rapid Refresh model are used as features to train these data-driven models. Training and evaluation are based on turbulence estimates of eddy dissipation rate (EDR) obtained from automated in situ aircraft reports. Our baseline RF model, consisting of 100 trees with 30 layers of maximum depth, significantly reduces forecast errors for $\text{EDR} < 0.1 \text{ m}^{2/3} \text{ s}^{-1}$ (which corresponds roughly to null and light turbulence) when compared with a simple regression model, increasing the probability of detection and in turn reducing the number of false alarms. Model complexity reduction via GBRT and feature-relevance analyses is performed, indicating that considerable execution speedups can be achieved while maintaining the model's predictive skill. Overall, the ML models exhibit enhanced performance in discriminating the EDR forecast among the light, moderate, and severe turbulence categories. In addition, these artificial intelligence techniques significantly simplify the generation of new NWP and grid-spacing specific turbulence forecast products.

KEYWORDS: Atmosphere; Turbulence; Forecasting techniques; Decision trees; Regression

1. Introduction

Turbulence events are a common threat to commercial and general aviation. While typically not fatal, turbulence encounters often result in serious injuries to aircraft occupants and/or rerouting of flights, significantly disrupting operations and reducing the capacity of the National Airspace System. To mitigate these adverse consequences associated with turbulence activity, turbulence forecasts are used by pilots, dispatchers, and air traffic controllers to strategically avoid regions of significant turbulence that can lead to injuries, aircraft fatigue, and damage (Sharman and Lane 2016). The forecasting difficulty is that these turbulence events are tied to scales that are smaller than those explicitly resolved by current operational numerical weather prediction (NWP) models. Therefore, specific methods are employed to downscale forecast grid-scale information (typically 1–10 km) in the form of turbulence diagnostics (or indices), which often involve spatial gradients, to aircraft-scale turbulence ($\sim 10\text{--}1000$ m).

One state-of-the-art turbulence forecasting system is the Graphical Turbulence Guidance (GTG) algorithm (Sharman and Pearson 2017; Muñoz-Esparza and Sharman 2018), which provides automated turbulence forecasts from regional to global coverage. Recent GTG developments have enabled a quantitative prediction of energy dissipation rate [$\text{EDR} \equiv \varepsilon^{1/3} (\text{m}^{2/3} \text{ s}^{-1})$, where ε is the turbulence dissipation], which is the standard for turbulence reporting by the International Civil

Aviation Organization (ICAO 2001), instead of the previously employed categorical qualitative aircraft weight dependent turbulence classes: “light,” “moderate,” and “severe” (see, e.g., Ellrod and Knapp 1992). Employing EDR as the forecast turbulence metric also has the advantage of being aircraft independent (Sharman et al. 2014). To provide an EDR forecast, the current GTG algorithm has three main components: (i) calculation of a suite of turbulence diagnostics, (ii) statistical mapping of these diagnostics into the EDR space, and (iii) determination of the best performing set of turbulence diagnostics that provide the GTG ensemble combination. The statistical linear mapping of the individual diagnostics to the EDR space is one of the key components of the current GTG algorithm. This transformation is accomplished by employing the method proposed by Sharman and Pearson (2017), which assumes a lognormal probability density function (PDF) of EDR at mid- and upper levels consistent with observed climatologies derived from in situ aircraft data (Lindborg 1999; Sharman et al. 2014; Kim et al. 2020). This statistical mapping technique has been extended by Muñoz-Esparza and Sharman (2018) to incorporate a dual distribution (lognormal and log-Weibull) to properly represent turbulence characteristics and account for the diurnal variability of turbulence at low levels (i.e., within the atmospheric boundary layer).

While the statistical mapping approach employed within the current GTG algorithm has demonstrated reasonable skill in producing EDR-based diagnostics that appropriately capture and forecast turbulence events, there are certain aspects of the current approach that can be potentially improved. One limitation of the current method is that turbulence diagnostics are assumed to have a lognormal distribution in a climatological sense (several years). In practice, we typically find only a few diagnostics that display a reasonable lognormal distribution

Denotes content that is immediately available upon publication as open access.

Corresponding author: Dr. Domingo Muñoz-Esparza, domingom@ucar.edu

DOI: 10.1175/JAMC-D-20-0116.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

FIG. 1. Example of PDFs of three turbulence indices (circles) and lognormal fits (solid lines). (a) Brown2 index (Brown 1973; $\text{m}^{2/3} \text{s}^{-1}$), (b) EDR derived from a tke-epsilon formulation by Marroquin (1998; $\text{m}^{2/3} \text{s}^{-1}$), and (c) a mountain wave turbulence-specific diagnostic from Sharman and Pearson (2017; $\text{K m}^2 \text{s}^{-2}$) using the 3D frontogenesis function combined with low-level flow characteristics. Solid circles indicate the selected bins included in the calculation of the best fit to a lognormal distribution. This figure is adapted from Sharman and Pearson (2017).

over the entire diagnostic's range. Figure 1 shows three example PDFs of turbulence diagnostics within the GTG ensemble taken from Sharman and Pearson (2017), derived from the 13-km Rapid Refresh (RAP; Benjamin et al. 2016) NWP model with a dynamical core based on the Weather Research and Forecasting (WRF) Model (Skamarock and Klemp 2008). An example of the replicated lognormal distribution over the entire range of diagnostic values is shown in Fig. 1a for the Brown2 (Brown 1973) index (based on a simplified form of the Richardson number tendency, converted to EDR). In contrast, there is often a departure from the climatological lognormal distribution, especially in the tails of the distribution, as seen in Figs. 1b and 1c for a tke-epsilon formulation by Marroquin (1998) and one of the mountain wave turbulence diagnostics of Sharman and Pearson (2017), respectively.

These types of distributions pose difficulties in fitting to a lognormal PDF required for the subsequent statistical mapping, with the fits becoming dependent on the range of values selected for the fitting and with forecast errors being amplified by the errors in the fit. Moreover, even if the climatology of a particular turbulence diagnostic does not exhibit a lognormal distribution, that does not necessarily imply the index does not have the potential to provide useful information for inferring turbulence levels. It is likely possible that a more complex mapping is required in these cases, or that a diagnostic is only useful under certain circumstances. Under the current GTG formulation, a turbulence diagnostic would be discarded if there is not at least some range of values that can be reasonably approximated by a lognormal distribution (typically toward the large magnitude end of the distribution). This decision is not a simple task and involves trial and error that requires significant human involvement.

Another component that could be improved in the current GTG algorithm is the selection of the diagnostics that constitute the final ensemble combination. The current method employs a forward-selection optimization technique that maximizes the skill of the ensemble prediction

for a given statistical metric of relevance. As the forecasting skill metric, the area under the receiving operating characteristic curve (AUC) is typically used, which represents the degree or measure of separability of positive and negative events (e.g., Gill 2016). At the end of the optimization process, a number of diagnostics are selected, which then get calculated and combined in a weighted average at each GTG evaluation (i.e., turbulence forecast). This method maximizes the overall statistical performance of GTG for the selected metric (AUC). However, it is expected that a more flexible algorithm, capable of selecting different turbulence diagnostics depending on specific atmospheric conditions, may result in more skillful turbulence forecasts.

Machine learning (ML) and artificial intelligence techniques provide an attractive alternative in pursuit of a more accurate turbulence forecast algorithm, given that they are capable of untangling complex patterns in data-driven models. ML techniques have recently experienced a boost in popularity and have been applied to a variety of meteorological problems (e.g., McGovern et al. 2017). Recent examples of the use ML for prediction of meteorological processes includes thunderstorm initiation (Williams et al. 2008), mesoscale convective system initiation (Ahijevych et al. 2016), solar irradiance (Gagne et al. 2017b), convective winds (Lagerquist et al. 2017), hail (Gagne et al. 2017a; Burke et al. 2020), 2-m temperature (Rasp and Lerch 2018), extreme precipitation (Herman and Schumacher 2018), storm longevity (McGovern et al. 2019a), wind power (Kosović et al. 2020), and severe weather (Hill et al. 2020), to name a few. In the context of turbulence forecasting, there have been a few early attempts to apply ML techniques to GTG-like framework since the early inception of the GTG algorithm.

Sharman et al. (2006) explored the applicability of several ML techniques to come up with an "optimal" combination of turbulence diagnostics that maximized the AUC. In particular, they compared three different models: logistic regression, tree classification, and neural networks, for the problem of turbulence

binary classification (i.e., null encounters of turbulence being 0 s, and moderate turbulence or greater encounters being 1 s). The authors found that when utilizing a reasonable amount of training data, the skill of these methods was comparable to the weighted ensemble of turbulence diagnostics as determined by maximizing the area under the ROC curves. Similar results were found by [Abernethy et al. \(2008\)](#) using logistic regression and support vector machines within GTG for categorical turbulence predictions (no turbulence, light, moderate and severe or greater). They obtained an AUC of 0.78 and 0.79 (1.0 being perfect), respectively, in comparison with an AUC of 0.78 based on the GTG method employed at that time. In a similar context, [Williams \(2014\)](#) applied the random forest technique (decision trees) for fusing data from diverse sources to produce a convective turbulence diagnosis, obtaining better skill with random forests than with the logistic regression technique.

The aim of the present work is to assess the ability of different machine learning techniques based on regression trees to develop an enhanced GTG-like algorithm that (i) overcomes some of the current weaknesses of the abovementioned optimization method, (ii) provides an alternative to the current mapping approach, (iii) results in an increased predictive skill relative to the additional complexity required, and (iv) is sufficiently computationally efficient to be incorporated in a real-time prediction system such as GTG in the near future. The remainder of the paper is organized as follows. [Section 2](#) describes the details of the data generation and training aspects of the ML models. Evaluation of the skill of the new ML-based approach versus a high-resolution GTG is presented in [section 3](#). Reduced-complexity ML methods are explored in [section 4](#). Insights into the more relevant physical processes for predicting turbulence from NWP model data as identified from the ML model are provided in [section 5](#). Some examples of the ML-based algorithm for specific cases are included in [section 6](#), with the conclusions and main findings of this study summarized in [section 7](#).

2. Data generation and training of the ML models

There are a wide variety of existing ML models that can be applied to the particular problem of interest, including logistic regression, support vector machines, regression trees, artificial neural networks, and other variants (e.g., [McGovern et al. 2017](#)). Herein we investigate the ability of regression trees (RTs) to perform turbulence predictions. We employ regression techniques, which allow us to produce a continuous variable such as EDR as the algorithm output (typically in the range $0.0\text{--}1.0\text{ m}^{2/3}\text{ s}^{-1}$), as opposed to the categorical output from decision trees. The regression-trees ML technique has a number of favorable characteristics that makes it amenable to the problem at hand. In particular, RTs are powerful algorithms capable of fitting complex datasets, and that are often used as a learning method in data mining while requiring less data preparation than other ML algorithms (e.g., neural networks). In a nutshell, RTs are built by progressively splitting the source dataset according to a given criteria, which depends on the particular range of a certain predictor variable (thresholding), resulting in the formation of branches and leaves.

To minimize excessive adaptation to the training dataset that often leads to a less generalizable model, known as overfitting, we employ the random forests (RFs) technique ([Breiman 2001](#)). Random forests is an ensemble ML technique based on RTs that introduces a random component when growing the trees during the training phase. This random component results in a greater tree diversity, hence reducing model overfitting. Instead of a single tree, now the outcome prediction of the ML algorithm is the ensemble mean of n_t trees, which become different by constraining the best splitting to a random subset of training features or predictors (bagging). While RFs will trade a higher bias for a lower variance, they generally yield an overall improved model. In addition, the individual predictions from the n_t trees in the forest can be used to derive a probability, which is a positive feature that we will explore in the future in the context of developing a probabilistic GTG product. Moreover, RFs have been found to be comparable to or even outperform other ML techniques for certain applications (e.g., [Lagerquist et al. 2017](#); [Gagne et al. 2019](#)).

The RF models were trained using a series of GTG-derived turbulence diagnostics prior to remapping into EDR space (i.e., raw turbulence diagnostics), as well as prognostic quantities from the NWP model forecast. As the underlying NWP model, we employ the WRF-based High-Resolution Rapid Refresh (HRRR) weather forecast product ([Smith et al. 2008](#)), with a horizontal grid spacing of 3 km.¹ Note this is a higher resolution than the current operational version of the 13-km RAP-based GTG described in [Sharman and Pearson \(2017\)](#). The higher-resolution HRRR-based GTG forecast product obtained employing the method by [Sharman and Pearson \(2017\)](#) was also derived as detailed at the end of this section, which will be termed HGTG hereinafter, and used as reference to evaluate the improvement in prediction skill by the RF models. To generate the training dataset, 16 months of retrospective HRRR Global Systems Division (GSD), version 2, files corresponding to the June 2018–September 2019 period were collected. We focused on the 6-h NWP model forecast lead time, including three initialization times, 0900, 1200, and 1500 UTC, with valid times that correspond to those hours over the United States containing the highest air traffic density ([Wolff and Sharman 2008](#)) but that are at the same time not too close together to contain correlated physical processes ([Sharman and Pearson 2017](#)). This resulted in a total of 1089 HRRR files, excluding the missing files due to data outages. The corresponding 1089 retrospective HGTG runs including all of the available turbulence diagnostics, raw variables from the NWP model output and scoring with in situ EDR and pilot reports (PIREPs) were efficiently generated on NCAR's supercomputer Cheyenne within a 24-h wall-clock time, performing simultaneous monthly runs parallelized on 36 cores using the message-passing interface (MPI) domain decomposition capability available

¹ The finest grid resolution for an operational NWP model that was publicly available over the contiguous United States at the time that this research was carried out.

within GTG. All of the ML models were trained and evaluated using Python's Scikit-learn library, version 0.20.2 (Pedregosa et al. 2011).

After having carried out the HGTG retrospective runs, a series of text files containing the scoring results, that is, pairing of GTG-derived raw turbulence diagnostics and NWP model prognostic fields with observations, were obtained. A post-processing code was developed to create a combined netCDF file from the independent HGTG runs that makes it possible to 1) select the altitude band of interest; 2) filter the data according to data quality flags, interpolation option, and type of observations; 3) remove unnecessary information related to the observations; and 4) select the type of GTG output for scoring (average, median, minimum, maximum, or nearest point). In the present study, we focus on the 20–45-kft (~6–14 km) altitude band, that is, cruising altitudes at upper levels, located in the upper troposphere and lower stratosphere (UTLS). No specific filtering is applied to isolate clear-air turbulence instances, and therefore we consider the best model that accounts for all sources of turbulence (also including mountain wave and convectively induced turbulence). Evaluations were based only on in situ EDR from United Airlines, Delta Air Lines, and Southwest Airlines. PIREPs were excluded given that with the actual dissemination of the in situ EDR algorithm, less than 8% of the observations corresponded to PIREPs, which in addition require a calibration to be converted to EDR (Sharman et al. 2014). Only data that had not been associated with any quality control issues were utilized. Similarly, interpolated data between true reports were discarded. As a result, a dataset containing 2.42 million model–observation pairs was created, including 114 predictors for each comparison instance, and organized as a 2D array for ease of manipulation within the ML modeling framework. The dataset was randomly shuffled and then split into independent training (75% of the total) and hold-out testing (25% of the total) subsets. In that way, we guarantee that the statistical distribution of the training and testing datasets is the same, giving no preference to certain time periods or turbulence classes.

Data-driven ML models typically have many configurable parameters (hyperparameters) that dictate the structure and complexity of the model. Among the random forests hyperparameters, we performed sensitivity analysis to the two most relevant ones: the number of trees $n_t = (20, 50, 75, 100, 300)$ and the maximum tree depth $t_d = (10, 30, 100)$. The minimum number of samples for splitting was also considered but is not discussed here given its correlation with the maximum tree depth. The rest of hyperparameters were used with their default values (Scikit-Learn-Developers 2018). Overall, it was found that reducing the tree depth results in less skill for low EDRs ($< 0.1 \text{ m}^{2/3} \text{ s}^{-1}$), with $t_d > 30$ yielding no noticeable improvements. Regarding the number of trees in the forest, $n_t = 100$ was found to provide good results, with more trees adding negligible skill to the model and a reduction of n_t producing a significant degradation of the model's skill. Therefore, for the following analysis we use the RF algorithm with 100 trees and maximum depth of 30 layers as our baseline model. This model was trained with the 114 available

predictors, consisting of raw NWP model output information, latitude, longitude, altitude, and turbulence diagnostics.

For the 3-km HRRR-based GTG turbulence forecasts utilized in this study, we employed a slightly different approach for selecting the turbulence diagnostics included in the final GTG ensemble. Sharman and Pearson (2017) determined the suite of indices conforming the GTG ensemble combination by maximizing the AUC. Instead, we used a multistatistic criterion, which combined the selected indices that optimized not only the AUC, but also the bias and true and Heidke skill scores, critical success index, mean absolute percentage error and several combinations of these metrics [AUC + true skill score, AUC/(1 + RMSE), (AUC + true skill score)/mean absolute percentage error (MAPE), and (AUC + true skill score)/(1 + volmog^{1/4}), where volmog is the fraction of the volume of predicted moderate or greater turbulence relative to the total predicted volume]. As a result, the following five turbulence diagnostics were selected for the ensemble GTG combination: absolute value of the total deformation squared (DEFSQ; Sharman and Pearson 2017), inertial advective wind (iawind; McCann 2001), divergence tendency divided by the local Richardson number (UBF/Ri; Sharman and Pearson 2017), averaged EDR derived from second-order structure functions of zonal and meridional velocities over the longitudinal direction (Frehlich and Sharman 2004), and the absolute value of the vorticity gradient times the gradient of the Montgomery streamfunction computed on isentropic surfaces (NCSU2/Ri; Sharman and Pearson 2017).

3. Evaluation of the baseline ML model and comparison to the HGTG

An initial overall assessment of the performance of the RF baseline model is provided by the two-dimensional histograms of in situ EDR versus the ML model and the HRRR-based GTG presented in Fig. 2. This plot is similar to a scatterplot, with the individual points substituted by probabilities over EDR bins (equally spaced in logarithmic space), and with the red line indicating a perfect one-to-one correlation. The baseline RF model exhibits certain features that differ from the current GTG approach. In particular, the RF algorithm displays a histogram distribution that is clearly clustered around the one-to-one perfect model across the entire EDR range. Although there is similar performance between the two methods in the range of higher EDR values (roughly equivalent to moderate-or-greater turbulence intensities), there is a noticeable improvement for $\text{EDR} < 0.1 \text{ m}^{2/3} \text{ s}^{-1}$ using the ML algorithm resulting in a significantly higher correlation with the observations. In contrast, the HGTG method exhibits a consistent overprediction of the observed EDR (increased probability of false detection), with a somewhat “flat” response for observed EDR values smaller than $0.1 \text{ m}^{2/3} \text{ s}^{-1}$, which is the suitable threshold for light turbulence for medium-sized aircraft at cruise altitudes (Sharman et al. 2014). This is partly due to the GTG optimization procedures that are scored based only on binary classification of “smooth” versus “moderate or greater” (taken as $\text{EDR} \geq 0.34 \text{ m}^{2/3} \text{ s}^{-1}$ for medium-weight category aircraft) turbulence categories. It is worth noting that the highest observed

FIG. 2. Two-dimensional histograms (percent of samples) of in situ EDR observations vs turbulence model forecasts for (a) the baseline RF algorithm ($n_t = 100$; $t_d = 30$) and (b) the HGTG approach. The red line indicates a perfect model (1:1 correlation). The histograms are based on a total number of 605 119 samples, corresponding to the testing data subset (25% of the total dataset). The lack of EDR data in some bins is due to different reporting strategies from some aircraft types.

probabilities are associated with $\text{EDR} < 0.1 \text{ m}^{2/3} \text{ s}^{-1}$, confirming that more intense turbulence encounters are a less frequent phenomenon than smooth conditions.

To quantify the skill of the developed RF algorithm for predicting EDR, several performance metrics were calculated. These include mean error (ME), mean absolute error (MAE), and MAPE, defined as follows:

$$\text{ME} = \frac{1}{n} \sum_{i=1}^n F_i - O_i, \quad (1)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |F_i - O_i|, \quad \text{and} \quad (2)$$

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^n \left| \frac{F_i - O_i}{O_i} \right|, \quad (3)$$

where n is the number of samples, and F_i and O_i are a given instance pair of forecast and observed EDR values, respectively.

It is often the practice in forecast assessment and verification to compute statistical metrics that are a single value, intended to be representative of overall algorithm performance. While this has the advantage of potentially simplifying the analysis, it can at the same time hide certain features of the model's skill. To be able to more thoroughly analyze the performance of the different turbulence prediction models, we present the distributions of ME, MAE, and MAPE grouped as a function of the observed EDR values (Fig. 3). This is specifically done to provide a quantitative assessment of the predictive skill of the models across the entire observed EDR range, thus allowing a more comprehensive understanding of model behavior. In addition, the distribution of samples in each of the observed EDR bins is included in the top panel of Fig. 3. There are several aspects that are noticeable. First, for $\text{EDR} < 0.1 \text{ m}^{2/3} \text{ s}^{-1}$, the developed RF model is significantly more accurate than the HGTG. This is in line with the 2D histogram distributions

shown in Fig. 2, and corroborated by a factor-of-3 smaller MAE ($\sim 0.027 \text{ m}^{2/3} \text{ s}^{-1}$) than the HGTG and a small positive bias (ME). For $\text{EDR} > 0.1 \text{ m}^{2/3} \text{ s}^{-1}$, both models present similar skill, with the HGTG slightly outperforming the ML model within the range $0.1 < \text{EDR} < 0.3 \text{ m}^{2/3} \text{ s}^{-1}$ (light-moderate turbulence). For severe turbulence, there is slightly better performance of the ML model, which in spite the reduced number of samples in that category ($n_{\text{obs}} = 613$ samples), is statistically significant as will be shown in the next section.

In the context of turbulence prediction, and given the predominance of nonturbulent conditions within the UTLS region, a very important aspect of a robust turbulence forecast model is to minimize false alarms (i.e., forecast turbulence activity during calm conditions). The ability of the RF and HGTG algorithms to forecast nonturbulent events is evaluated in Fig. 4, where the probability of predicted EDR during calm conditions, $\text{EDR} = 0.01 \text{ m}^{2/3} \text{ s}^{-1}$, is presented. These instances correspond to the $[0.0, 0.02) \text{ m}^{2/3} \text{ s}^{-1}$ range, truncated by the in situ EDR algorithm to report a value of $0.0 \text{ m}^{2/3} \text{ s}^{-1}$, and that we have presented at the center of the bin in the previous analysis. In addition to the distribution of forecast EDR, a cumulative PDF is included (solid lines). A perfect model would have all instances at the lowest EDR bin, and in turn, a cumulative distribution with a rapid growth as EDR increases (similar to a ROC curve). It is clear how the RF model has a distribution that is considerably skewed toward lower EDR values relative to the HGTG. One-half of the times that calm conditions are observed, the RF algorithm forecasts $\text{EDR} < 0.02 \text{ m}^{2/3} \text{ s}^{-1}$, and in 90% of the cases its predictions are $< 0.1 \text{ m}^{2/3} \text{ s}^{-1}$ (i.e., below the typical light turbulence threshold). This frequency of occurrence is more consistent with the observed EDR climatology in Sharman et al. (2014). In contrast, the HGTG algorithm results in an $\text{EDR} > 0.1 \text{ m}^{2/3} \text{ s}^{-1}$ about 50% of the time, indicating a significantly higher false-alarm

FIG. 3. Mean error distributions (ME, MAE, and MAPE) as a function of observed EDR for the RF and HGTG algorithms. The solid lines represent the mean plus 1 std dev. Also shown is the number of samples in each observed EDR bin. The overlapping part of the two compared histograms is colored in gray. The histograms are based on the testing data subset, corresponding to a total number of 605 119 samples (25% of the total dataset).

rate (FAR; i.e., predicting some level of turbulence activity in calm conditions) relative to RF.

4. Regression-tree ML models of reduced complexity

The present results obtained with the baseline RF model are very encouraging as a way to improve certain aspects of turbulence predictions within the GTG or other ensemble-based algorithms. With an end goal of implementing the developed ML model for operational use, we assessed the computational requirements associated with this more complex model. For the baseline RF model (100 trees with a maximum depth of 30 layers), it would take ~ 30 s to produce a full HRRR grid ($1799 \times 1095 \times 50$) of HGTG forecasts (after indices are computed) assuming similar computational resources for parallel MPI execution to those available within the current operational setup (i.e., utilizing 96 cores). The parallelized GTG algorithm running on 96 cores only requires ~ 30 s to run a HRRR case from beginning to end, essentially doubling the execution time with the respect to the HGTG algorithm (note that the RF models require prior execution of GTG to derive raw turbulence diagnostics used as features by these ML models). While these execution times would still allow a seamless operational implementation of an ML-based GTG-like algorithm under the current execution time requirements, challenges could arise when the underlying NWP models have larger grids, as it is often the case with global models.

To minimize the complexity of the ML model, and in turn the associated execution time, we explored two different

options. The first one is the use of another variant of regression trees, called gradient-boosted regression (GBRT; Friedman 2001). GBRT is also an ensemble technique, but it is slightly different from the RF technique. Instead of utilizing a random subset of the available predictors to obtain

FIG. 4. PDFs of EDR forecasts for the observed $[0.0, 0.02) \text{ m}^2/3 \text{ s}^{-1}$ bin from the RF and HGTG algorithms. The solid lines are the cumulative PDFs, integrated from low to high EDR values. The overlapping part of the two compared histograms is colored in gray. The histograms and cumulative PDFs are based on a total number of 349 848 samples (57.8% of the testing data subset).

FIG. 5. Two-dimensional histograms (percent of samples) of in situ EDR observations vs turbulence model forecast for the GBRT model with $n_t = 60$ trees and maximum $t_d = 10$ layers. The overlapping part of the two compared histograms is colored in gray. The histograms are based on a total number of 605 119 samples, corresponding to the testing data subset (25% of the total dataset), comparing the GBRT with the baseline RF model.

the optimum splitting strategy, in the case of GBRT, the first tree targets the actual EDR observations as the predictand, while each subsequent tree minimizes the residual obtained after all of the previous trees (i.e., the remaining error). This approach has the caveat of being more susceptible to overfitting but allows application of early stopping. This means, in practice, that the optimum number of trees can be determined, at which point further increase in the number of trees no longer provides any skill improvement (this is not as straightforward with RFs).

By using the GBRT technique, we were able to simplify the ML model to $n_t = 60$ trees and $t_d = 10$ layers of maximum depth. This results in a reduction of the model complexity that in turn cuts the execution time on a HRRR-like grid from 30 s down to 8 s (a factor of 3.75). The 2D histograms of in situ EDR versus model predictions are presented in Fig. 5, comparing the baseline RF and the GBRT algorithm error distributions. The GBRT model behavior is qualitatively similar to that of the RF algorithm. There is, however, an increase in the model's spread from the perfect correlation (red line), partly attributed to the reduced complexity of the model and to the more susceptible nature of the GBRT models to overfit the training dataset. This results in a small increase of model error with respect to the baseline RF model, as seen from the ME, MAE, and MAPE distributions (Fig. 5, right panel). Nevertheless, it is worth noting that a model of the same complexity (same n_t and t_d as the GBRT model) led in the case of the RF technique to lower predictive skill (not shown). Therefore, an ML model employing the GBRT technique is a viable option if operational execution time requirements pose a constrain to the complexity of the ML algorithm.

Another advantageous quality of random forests is that they allow determination of the relative importance of the predictors in a straightforward manner. After training the RF model, one can calculate how much each feature reduces the impurity

(a measure of how well the trees split the data) for each node of the tree and for each tree of the forest algorithm (Breiman 2001; Louppe et al. 2013). In this study we use the mean squared error as splitting criterion, which is equal to variance reduction. This metric enables ranking the predictors according to their contribution to the accuracy of the forecast EDR. Exploiting this feature of RFs, we developed an algorithm to perform a second training phase after feature importances have been determined in a previous training phase, and that allows selection of the degree of model predictor total relevance to retain. This algorithm excludes the least relevant predictors up to the specified threshold for total relevance and performs a new training with the reduced set of indices.

The error distributions for a retained 60% and 15% feature-relevance thresholds are shown in Figs. 6a and 6b, labeled RFri60 and RFri15, respectively. As it can be seen, decreasing the relevance level down to 60% does not have a large impact in the resulting RF model. This is a significant reduction in the number of predictors, from 114 to 32, which, while it does not impact the ML model execution directly (i.e., it utilizes the same n_t and t_d values), it does indirectly speed up the overall execution of the GTG-like algorithm through a smaller number of turbulence indices that need to be computed. The error distributions in Fig. 6a display a slight improvement of RFri60 with respect to the baseline RF model for $EDR < 0.1 \text{ m}^{2/3} \text{ s}^{-1}$, indicating that the additional 82 features employed by the latter were mostly contributing to the model overfit. Indirectly, the RF model complexity reduction via feature-relevance-importance analysis helps to identify the predictors that are highly correlated, eliminating them before the model skill starts to be reduced. As the relevance level is further reduced to 15%, only utilizing 2 features, the skill of the predictions starts to degrade for $EDR > 0.1 \text{ m}^{2/3} \text{ s}^{-1}$ (see Fig. 6b). However, it is worth emphasizing the remarkable performance of the RFri15 model, which only utilizes the wind speed at the

FIG. 6. 2D histograms (percent of samples) of in situ EDR observations vs turbulence model forecast for the reduced RF models for (a) relative importance based on impurity of 60% (32 features) RFRi60, (b) relative importance based on impurity of 15% (2 features) RFRi15, and (c) relative importance based on permutation of 89% (32 features) RFRp89, all using the same number of trees and maximum depth as the RF baseline model ($n_t = 100$, $t_d = 30$). The red line indicates a perfect model (1:1 correlation). The overlapping part of the two compared histograms in each panel is colored in gray. The histograms are based on a total number of 605 119 samples, corresponding to the testing data subset (25% of the total dataset), comparing each of the reduced-complexity models with the baseline RF model.

tropopause and the potential vorticity gradient as predictors, evidencing the advantage of complex ML-based models. At the same time, the lower errors by the RFri15 model for $0.01 < \text{EDR} < 0.1 \text{ m}^{2/3} \text{ s}^{-1}$ indicate that predicting calm conditions is the more straightforward task, and that the real challenge resides on predicting the more rare turbulence episodes.

Impurity importance has the advantage of being easy to compute, and it is therefore a popular approach in interpreting the relevance of predictors in RT-based ML methods. However, it is derived from the resulting model, and therefore suffers from only taking into account statistics from the training phase/dataset. This can potentially lead to certain features being identified as relevant as long as the ML model has the ability to use them to overfit. Feature permutation is another interpretation method that is useful in determining the relative importance of the different predictors in the RF model, and that can mitigate this limitation. The feature permutation technique measures importance by sequentially permuting each feature with other predictors in the model and then evaluating the increase in the error (either using the training or the hold-out testing datasets). If performance deteriorates significantly when permuting a feature, this indicates that the feature is important. We performed permutation importance analysis applying 30 permutations for each of the 114 available predictors. The results from the reduced-complexity model of 32 features, and that retains 89% of the total relevance as calculated from permutation importance on the testing dataset, RFrp89, are also included in Fig. 6c. Both the scatterplot and error distributions are very similar to the ones from RFri60, which indicates that the reduced models are generalizable (i.e., not subject to overfitting), since impurity is based upon training while permutation relevance was calculated using the hold-out testing dataset. Moreover, RFri60 and RFrp89 share 72% of their features, which are further discussed in section 5, pointing to the robustness of the feature selection approach for these reduced-complexity RF models.

Up to now, model skill analysis has mainly concentrated on the inspection of ME, MAE, and MAPE distributions for the different algorithms. To provide an assessment consistent with previous works, we calculate additional statistical metrics commonly employed in the literature to evaluate meteorological forecasts: true skill score (TSS), probability of detection or hit rate (PODY), probability of detecting out-of-class events (PODN), and AUC, calculated from 2×2 contingency tables (e.g., Jolliffe and Stephenson 2012; Gill 2014). Note that here we separate the errors as a function of the observed EDR as follows: “null” ($0.0\text{--}0.05 \text{ m}^{2/3} \text{ s}^{-1}$), “very light” ($0.05\text{--}0.15 \text{ m}^{2/3} \text{ s}^{-1}$), “light” ($0.15\text{--}0.22 \text{ m}^{2/3} \text{ s}^{-1}$), “moderate” ($0.22\text{--}0.34 \text{ m}^{2/3} \text{ s}^{-1}$), and “severe” ($0.34\text{--}1.00 \text{ m}^{2/3} \text{ s}^{-1}$) to account for the differences in model performance upon the EDR range, which otherwise would be hidden if a single metric were computed.

As shown in Table 1, the RFri60 and RFrp89 models exhibit the lowest errors (ME, MAE, and MAPE) and a large hit rate of 0.72 for null and very light turbulence, closely followed by RF and GBRT models. This indicates good skill by the ML-based models in predicting smooth conditions, consistent with the results presented in Fig. 4. In contrast, HGTG has a very low PODY of 0.05, related to the overforecasting trend

observed in the 2D histograms from Fig. 2b for low EDR values (i.e., hot model response), and with a MAPE of 367% (3.3 times as large as RFri60 and RFrp89, with a MAPE of 112%). While the HGTG has the highest PODN (almost 1.0), this is a consequence of systematically placing forecast EDR values outside the null category. For a more consistent evaluation, the TSS should be considered, since it combines the correctness of the model in predicting values inside the selected range together with the skill in forecasting values not observed within the particular EDR range. The RFrp89 and RFri60 models have a TSS of 0.55 for null conditions, while the HGTG has a considerably lower value (0.042). This further confirms that the increased probability of detecting values outside the null range comes at the expense of a reduced hit rate.

A change in the predictive skill of the different models starts to emerge for $\text{EDR} > 0.15 \text{ m}^{2/3} \text{ s}^{-1}$. All of the models perform similarly for light and moderate turbulence, with the MEs having in all instances a negative sign, indicative of a trend to under predict turbulence intensity. The HGTG has the lowest errors, with a MAPE that is 3.5% smaller than that of the RF model for light turbulence and 0.9% smaller for moderate turbulence (with slightly increased errors for RFrp89, RFri60, and GBRT). In contrast, the ML models have slightly larger TSS, PODY, and AUC. For severe turbulence, there is a marginal improvement by the RF, RFrp89, and RFri60 models, with a 5.0% reduction in MAPE compared to HGTG, as well as the highest TSS, PODY, PODN, and AUC and the lowest FAR. In particular, the ML models consistently lead to higher values of area under the ROC curve, employing as threshold the middle of the EDR range for each of the turbulence categories. The largest AUC values are found for all models for the moderate turbulence regime, with an AUC of 0.91 for the RFrp89 and RFri60 models and an AUC of 0.79 by the HGTG. In general, both the HGTG and ML models have increased MAPE by $\sim 15\%$ for severe turbulence compared to the moderate turbulence category, with similar PODY values, indicating that the spread of the model is larger when EDR is forecast outside the severe turbulence range. We attribute this to the difficulties in predicting complex highly transient severe events, which are rarely observed (0.1% of time), and for which it is difficult to skillfully train any model (see the drop off in frequency of observation at the top panel of Fig. 3 for $\text{EDR} > 0.3 \text{ m}^{2/3} \text{ s}^{-1}$).

Note that the model evaluation scores presented in this section are statistically significant. To support this statement, we performed a k -fold cross-validation analysis, which allows to estimate the robustness of the developed models subject to a limited training dataset. Given the 75/25 split of the data we utilized, a fourfold approach was implemented. Table 2 displays the variability in the statistical metrics for each of the turbulence categories using the RF baseline model when different data partitions are used to train the same ML model (i.e., no overlapping in the testing datasets among the different folds). The maximum variability among the four trained models is expressed both in absolute range σ_{max} and in relative percentage to the mean of the four RF models σ_{rel} for a better interpretability of the variability ranges.

TABLE 1. Statistical scores for the baseline RF, the GBRT, and the RFri60 and RFrp89 models relative to the current GTG, including mean error, mean absolute error, mean absolute percentage error, true skill score, probability of detection, probability of correct negatives, false-alarm rate, and area under the curve. The observed EDR range is split into relevant turbulence categories: null, very light, light, moderate, and severe. The threshold for AUC calculation corresponds to the middle of the EDR interval for each category. The statistical metrics are based on the testing data subset, corresponding to a total number of 605 119 samples (25% of the total dataset). The best-performing model for each metric within a turbulence category is highlighted in boldface font, with uncertainty bounds calculated for each metric within a specific turbulence category on the basis of the fourfold cross-validation analysis presented in Table 2, below.

Algorithm	ME ($\text{m}^{2/3} \text{s}^{-1}$)	MAE ($\text{m}^{2/3} \text{s}^{-1}$)	MAPE (%)	TSS	PODY	PODN	FAR	AUC
Null: $0.0 \leq \text{EDR}_{\text{insitu}} < 0.05 \text{ m}^{2/3} \text{s}^{-1}$ ($n_{\text{obs}} = 482\,893$ samples)								
RF	0.030	0.032	134.531	0.531	0.663	0.867	0.048	0.797
GBRT	0.028	0.034	140.165	0.452	0.622	0.829	0.065	0.776
RFri60	0.024	0.027	112.063	0.553	0.721	0.832	0.056	0.796
RFrp89	0.024	0.027	112.399	0.557	0.720	0.837	0.054	0.798
HGTG	0.084	0.084	367.332	0.042	0.050	0.992	0.039	0.674
Very light: $0.05 \leq \text{EDR}_{\text{insitu}} < 0.15 \text{ m}^{2/3} \text{s}^{-1}$ ($n_{\text{obs}} = 98\,300$ samples)								
RF	0.028	0.049	54.652	0.264	0.583	0.681	0.738	0.885
GBRT	0.021	0.051	56.930	0.276	0.576	0.701	0.728	0.846
RFri60	0.018	0.045	49.744	0.327	0.599	0.727	0.701	0.889
RFrp89	0.019	0.045	49.866	0.328	0.602	0.726	0.701	0.890
HGTG	0.047	0.057	67.161	−0.145	0.675	0.181	0.862	0.731
Light: $0.15 \leq \text{EDR}_{\text{insitu}} < 0.22 \text{ m}^{2/3} \text{s}^{-1}$ ($n_{\text{obs}} = 15\,654$ samples)								
RF	−0.021	0.067	34.589	0.256	0.314	0.942	0.875	0.875
GBRT	−0.040	0.076	39.135	0.197	0.252	0.945	0.891	0.837
RFri60	−0.029	0.071	36.564	0.232	0.277	0.956	0.858	0.878
RFrp89	−0.028	0.070	36.331	0.234	0.278	0.956	0.857	0.879
HGTG	−0.035	0.060	31.120	0.171	0.294	0.877	0.940	0.734
Moderate: $0.22 \leq \text{EDR}_{\text{insitu}} < 0.34 \text{ m}^{2/3} \text{s}^{-1}$ ($n_{\text{obs}} = 7659$ samples)								
RF	−0.071	0.109	41.539	0.188	0.200	0.988	0.828	0.903
GBRT	−0.094	0.123	46.715	0.132	0.146	0.985	0.887	0.868
RFri60	−0.076	0.115	43.692	0.179	0.189	0.990	0.803	0.905
RFrp89	−0.076	0.114	43.498	0.182	0.192	0.990	0.807	0.906
HGTG	−0.100	0.107	40.636	0.094	0.126	0.968	0.952	0.788
Severe: $0.34 \leq \text{EDR}_{\text{insitu}} < 1.00 \text{ m}^{2/3} \text{s}^{-1}$ ($n_{\text{obs}} = 613$ samples)								
RF	−0.174	0.218	50.187	0.175	0.178	0.998	0.931	0.785
GBRT	−0.194	0.241	55.728	0.162	0.165	0.997	0.951	0.770
RFri60	−0.177	0.220	50.779	0.182	0.184	0.998	0.928	0.813
RFrp89	−0.175	0.219	50.507	0.190	0.192	0.998	0.924	0.835
HGTG	−0.241	0.241	55.451	0.029	0.031	0.998	0.982	0.696

The fourfold cross-validation results summarized in Table 2 clearly show that our data splitting approach provides statistically significant results, with the maximum variability being typically lower than a few percent. This is in part due to the considerably large dataset utilized (2.42 million samples) and the random split between training and testing datasets, which ensures the data distribution is the same among the two datasets. The severe turbulence category exhibits an increase in variability of the different metrics with respect to the lower EDR categories, an expected outcome due to the considerably smaller number of samples ($\approx 0.1\%$ of the occurrences in the dataset), but still remains reduced. The σ_{max} values were used as a proxy for uncertainty ranges in the skill of the developed models, enabling identification of the differences that are truly significant from those attributed to insufficient training data for the ML models. Therefore, when the decrease in skill from the best performing model is smaller or equal to σ_{max} , several models are highlighted as best performing for a given

turbulence category and statistical score, indicative of differences among models being smaller than their statistical significance.

5. Physical interpretation of the ML model

Despite the widespread use of ML techniques and their demonstrated benefit within the meteorological community, these machine learning approaches are often criticized for being “black boxes” that lack physical interpretability (e.g., McGovern et al. 2019b). The impurity and permutation importance methods, applied in the previous section for reducing the complexity of the RF algorithm, can also be employed to identify what the relevant features of the model are, and relate them to the corresponding turbulence mechanisms in the atmosphere. For the regression tree methods employed herein, feature impurity reveals the predictors that affect the most training samples (i.e., higher in tree) and that split the data more effectively (i.e., decrease impurity more), while the

TABLE 2. Variability of the statistical scores from the k -fold cross-validation analysis for the baseline RF model. A fourfold approach was employed ($k = 4$), to be consistent with the partitioning between training (75%) and testing (25%) datasets. For each metric, the maximum difference among the fourfold $\sigma_{\max}$ and the relative percentage of that difference with respect to the mean of each turbulence category σ_{rel} are presented. The statistical metrics are based on the testing data subset, corresponding to a total number of 605 119 samples (25% of the total dataset). The $\sigma_{\max}$ values are used in Table 1 to account for statistical significance in identifying the best skill model.

Algorithm	ME ($\text{m}^{2/3} \text{s}^{-1}$)	MAE ($\text{m}^{2/3} \text{s}^{-1}$)	MAPE (%)	TSS	PODY	PODN	FAR	AUC
Null: $0.0 \leq \text{EDR}_{\text{insitu}} < 0.05 \text{ m}^{2/3} \text{s}^{-1}$ ($n_{\text{obs}} = 482\,304\text{--}483\,141$ samples)								
$\sigma_{\max}$	0.000	0.000	0.400	0.004	0.001	0.003	0.001	0.001
σ_{rel} (%)	0.000	0.000	0.297	0.757	0.151	0.347	2.051	0.126
Very light: $0.05 \leq \text{EDR}_{\text{insitu}} < 0.15 \text{ m}^{2/3} \text{s}^{-1}$ ($n_{\text{obs}} = 98\,300\text{--}98\,696$ samples)								
$\sigma_{\max}$	0.001	0.000	0.264	0.006	0.007	0.001	0.002	0.002
σ_{rel} (%)	3.670	0.000	0.484	2.284	1.204	0.147	0.271	0.226
Light: $0.15 \leq \text{EDR}_{\text{insitu}} < 0.22 \text{ m}^{2/3} \text{s}^{-1}$ ($n_{\text{obs}} = 15\,654\text{--}16\,050$ samples)								
$\sigma_{\max}$	0.004	0.001	0.483	0.005	0.005	0.001	0.004	0.013
σ_{rel} (%)	−4.819	1.487	1.017	0.392	0.000	0.106	0.343	0.343
Moderate: $0.22 \leq \text{EDR}_{\text{insitu}} < 0.34 \text{ m}^{2/3} \text{s}^{-1}$ ($n_{\text{obs}} = 7486\text{--}7659$ samples)								
$\sigma_{\max}$	0.001	0.002	0.353	0.001	0.000	0.001	0.004	0.003
σ_{rel} (%)	−5.594	1.840	1.162	2.639	2.472	0.101	0.643	1.442
Severe: $0.34 \leq \text{EDR}_{\text{insitu}} < 1.00 \text{ m}^{2/3} \text{s}^{-1}$ ($n_{\text{obs}} = 550\text{--}632$ samples)								
$\sigma_{\max}$	0.009	0.008	1.381	0.028	0.025	0.000	0.020	0.082
σ_{rel} (%)	−5.248	3.747	2.778	14.698	14.489	0.000	2.160	10.599

feature permutation on the testing dataset identifies the most relevant diagnostics as the ones that result in larger errors when compared to the original model as their features are permuted with other predictors.

Inspection of the 10 most relevant features used by the RFri60 and RFrp89 models, collected in Table 3 and ranked by relative importance for each criterion, exposes several interesting characteristics of these ML models. Among the 33 features used by the RFri60 model, there are 11 that are diagnostics related to the state and properties of the atmosphere surrounding the tropopause (one-third of the total features). In particular, the feature with the largest relative importance is the maximum vertical wind shear in the vicinity of the tropopause (TROPVWS), with a relative importance of 12.2% based on impurity. Moreover, all of the tropopause-based indices have a combined relevance of 37.6%, indicating that tropopause dynamics play a significant role in the occurrence of turbulence events observed at cruise altitudes. This finding is consistent with the results from the permutation importance analysis, which also select TROPVWS as the most relevant diagnostic, and with a combined relevance of 35.79% for the tropopause related diagnostics (10 features). Abundant empirical evidence exists for higher levels of turbulence near the tropopause region (e.g., Partl 1962; Chandler 1987; Lester 1993; Wolff and Sharman 2008). These turbulence-conductive phenomena include jet streams and fronts linked to tropopause folds, which are often associated with turbulence in the UTLS (e.g., Shapiro 1978, 1980; Koch et al. 2005; Sharman et al. 2012; Trier et al. 2020). Not surprisingly then, the feature with the second largest relevance is the horizontal potential vorticity gradient (PVGRAD), with a relative importance of 4.8%. This is likely attributed to the relation between gradients of potential vorticity and stratospheric air intrusions at lower levels

(tropopause folds), but also indicative of strong vertical shear that can lead to turbulence production (e.g., Danielsen 1968).

Another notable correlation found by the ML model between the features and the predictand, that is, turbulence in the form of EDR, is that of moisture-related variables. This is also not an unexpected outcome, given the well accepted interplay between moist convection and turbulence events in the UTLS region (e.g., Sharman et al. 2012; Sharman and Trier 2019), that is, the model is skillful at predicting convective-induced turbulence events (provided the convection is properly captured by the underlying NWP model). Similarly, the mean mixed-layer mixing ratio (MLMR), corresponding to the first 500 m above the surface, is ranked as the second most relevant feature by the permutation importance criterion, with a relative importance of 12.8%, and as the third most relevant feature by the impurity analysis, with a relative importance of 4.0%. We attribute this to the requirement of a moisture source at the surface for moist convection to develop at higher altitudes. Interestingly, the mean mixed-layer potential temperature is also one of the 32 model features, with relative impurity and permutation importances of 2.4% and 1.11%, and that can be interpreted as an indicator of the strength for air parcels to be buoyantly lifted from the boundary layer into the free troposphere (the higher is the likelihood of being warmer than the air aloft). In addition, the water vapor mixing ratio q_v is ranked as the third most relevant parameter when considering permutation (relative importance of 12.8%) and fifth for impurity (relative importance of 3.9%), indicative of the additional importance of the vertical distribution of moisture throughout the atmosphere.

The two relevance identification methods also recognize the importance of the small scales explicitly captured by the NWP model, by including the resolved turbulence kinetic

TABLE 3. List of the top 10 features and their description, ranked by decreasing relative importance (in parentheses) as identified by the impurity (RFri60 model) and the permutation (RFrp89 model) importances for reduced models of 32 features.

Feature	Description	Unit	Impurity (RFri60)	Permutation (RFrp89)
The top 10 features				
TROPVWS	Max vertical shear in the vicinity of the tropopause	s^{-1}	12.245 (1)	22.529 (1)
PVGRAD	Horizontal potential vorticity gradient	$PVU\ km^{-1}$	4.848 (2)	7.183 (5)
MLMR	Mean mixed-layer mixing ratio	$g\ kg^{-1}$	4.020 (3)	12.828 (2)
RTKE	Resolved turbulence kinetic energy	$m^2\ s^{-2}$	3.996 (4)	9.201 (4)
q_v	Water vapor mixing ratio	kg^{-1}	3.986 (5)	9.266 (3)
z	Altitude	mj	3.690 (6)	4.103 (6)
lon	Longitude	°	3.270 (7)	3.172 (7)
lat	Latitude	°	3.239 (8)	1.925 (12)
TROPUMAX	Distance between max wind speed above/below the tropopause	m	2.949 (9)	2.328 (10)
w^2/Ri	Square of the vertical velocity divided by the Richardson number	$m^2\ s^{-2}$	2.854 (10)	3.066 (9)
TROPSTAB	Tropopause height based on stability change	m	2.451 (24)	3.164 (8)
Different features				
divm	Absolute value of horizontal divergence	s^{-1}	2.391 (26)	—
LMO	Obukhov length	m	2.513 (19)	—
EDRRCH	EDR derived from a Richardson number tendency (Roach 1970)	$m^{2/3}\ s^{-1}$	2.513 (27)	—
UMAXT	Max wind within the atmospheric boundary layer	$m\ s^{-1}$	2.513 (23)	—
DRAGYT	Wave drag in the meridional direction	$N\ m^{-2}$	2.270 (31)	—
Endlich	Wind speed times turning (Endlich 1964)	$radian\ s^{-1}$	2.453 (22)	—
UBF/Ri	Divergence tendency	s^{-2}	—	1.175 (17)
1/RiTW	Inverse Ri based on shear derived from the thermal wind equations	—	—	0.824 (25)
NGM1	Wind speed times deformation (Reap 1996)	$m\ s^{-2}$	—	0.820 (26)
TROPWMO	Tropopause height from WMO definition	m	—	0.712 (27)
w^2	Vertical velocity squared	$m^2\ s^{-2}$	—	0.696 (28)
SGSTKE/Ri	Subgrid TKE from NWP model	$m^2\ s^{-2}$	—	0.641 (32)

energy calculated as the deviation from a local spatial average (RTKE), as the fourth-ranked relative importance feature in both cases (9.2% from permutation and 3.9% from impurity). In addition to the RTKE, there are several other indices that relate to resolved small-scale dynamics and that are part of the suite of the ML model's features, including vertical velocity squared over the Richardson number, horizontal divergence, vertical velocity variance, and a dissipation rate derived from a Richardson number tendency. This suggests that high-resolution NWP models provide useful information that can be downscaled by the ML algorithm and mapped to an aircraft-scale EDR estimate. It is worth mentioning that altitude, longitude and latitude are among the highest rated features, with a combined relative importance of $\approx 10\%$ in both importance metrics, indicating that the ML model is able to identify spatial patterns associated with turbulence occurrences. The importance of these localization-enabling features is of specific relevance for classifying turbulence generation mechanisms under a given spatially evolving weather scenario, in addition to local terrain features associated with influencing weather systems and associated with generation events in the form of mountain waves.

Moreover, the same nine features are in both of the top-10 rankings of importance and have similar relative positions. Of the 32 predictors, only 6 features differ between the two reduced models, listed in the bottom part of Table 3. This distribution of importance across multiple features provides further confidence in the robustness and generalizability of the developed reduced models. A further attempt to elucidate how each of these relevant features affects the resulting ML model response was made by deriving partial-dependence plots, which calculate the average EDR forecast for each possible value of a feature. While a few features like TROPVWS, q_v , and MLMR have partial-dependence plots with regions of positive slope, overall, it is found that the majority of the features result in a flat partial-dependence response over a large portion of its range (even in the reduced models), which indicates that our model often predicts EDR based on complex nonlinear interactions that involve multiple predictors.

6. Examples of the ML-based algorithm

Two specific cases are used to illustrate the representative behavior of the developed RF algorithms. These are by no means examples of the best forecasts that can be obtained with

FIG. 7. Examples of EDR turbulence forecasts ($\text{m}^{2/3}\text{s}^{-1}$) with the RFri60 model for (left) a CAT event and (right) a CIT case. Contours of ML-based RFri60 model EDR FL320 6-h predictions valid at (a) 1800 UTC 21 Jun 2019 and (d) 1500 UTC 18 May 2019 are shown. (b),(e) Observations that include $\pm 300\text{-ft}$ ($\sim 90\text{ m}$) altitude and $\pm 10\text{-min}$ time interval from the specific forecast and observed composite radar reflectivity (dB); EDR observations from automated in situ sensors are indicated by crosses (circles show null and very light turbulence). (c),(f) Scatterplots of observed-to-forecast EDR, where the black solid line represents a perfect model (1:1 correlation), and the two dashed lines indicate deviations by a factor of 2. Scatterplot forecast values are taken as the maximum within a 45-km-per-side square box, to account for spatial shifts in the turbulence patterns derived from the NWP model with respect to reality.

our ML-based turbulence prediction model, rather an attempt to provide examples of typical EDR forecasts from the RF baseline model. Also, it is worth reminding that NWP model errors often propagate into our turbulence forecast algorithm degrading its skill, and that there is no simple approach to separate NWP model forecast errors from errors in the actual turbulence forecast algorithm, already inherited during the training phase (whether it uses ML techniques or any other type of methods).

The first case corresponds to a clear-air turbulence (CAT) episode, which developed toward the East Coast, approximately at the boundary of West Virginia with Virginia on 21 June 2019 (Fig. 7b). The RFri60 model 6-h EDR forecast from Fig. 7a displays a localized patch of enhanced turbulence activity in the same region, which extends farther north, and that agrees well with the reported in situ EDR in timing, location, and intensity. There is another region of light-to-moderate CAT over northern Wisconsin and Minnesota,

which is also well predicted by the RFri60 model. The convective core over Missouri and Illinois, as indicated by the observed radar reflectivity, is identified by the model as a region of moderate turbulence; however, there are no reports available to evaluate its accuracy. The ML-based turbulence forecast further exhibits good skill in identifying the regions over central and western United States where null turbulence was reported, as depicted by a white EDR contour level. The scatterplot for this event in Fig. 7c depicts an accurate response of the model, locating the majority of the observation-to-forecast pairs within a factor of 2 for observations within the light and moderate turbulence categories, with a high PODY of 0.89 and reduced MAE and MAPE of $0.057 \text{ m}^{2/3} \text{ s}^{-1}$ and 28.8%, respectively, as well as for the five severe turbulence reports, which have a MAPE of 16.7%.

The other example that we examine is a case occurring on 18 May 2019 where near-cloud turbulence was reported. The observed mosaic maximum reflectivity valid at 1500 UTC (Fig. 7e) shows a coherent squall line over central Texas and Oklahoma. The overlaid turbulence reports display a number of light-to-moderate (37) and a few severe (4) reports within the convectively active area. The EDR forecast from the RF model in Fig. 7d, corresponding to a 6-h forecast valid at 1500 UTC, predicts elevated turbulence within a region that reasonably matches the location of the active convection not only over the central United States but also toward the northern United States and West Coast, as indicated by the maximum radar reflectivity contours. Again, the majority of the observed light-to-moderate turbulence instances ($0.1 \leq \text{EDR} \leq 0.34 \text{ m}^{2/3} \text{ s}^{-1}$) are clustered within a factor of 2 of the observations, with a very high PODY of 0.909 and a MAE of $0.05 \text{ m}^{2/3} \text{ s}^{-1}$ (representing 14%–50% of the spanned EDR range, MAPE = 30.8%). An interesting feature of the ML model is that it is able to predict a sharp line of high turbulence ($\text{EDR} \sim 0.5 \text{ m}^{2/3} \text{ s}^{-1}$) that matches the deep core of the squall line as indicated by the observed mosaic radar reflectivities, especially over Texas. This is also a very encouraging finding, given that in-cloud turbulence is most often avoided by pilots, since it is detectable with radars, and therefore the number of in-cloud events used for training of the algorithm is typically sparse.

7. Summary and conclusions

We investigate the applicability of ML techniques, in particular regression-tree-based methods, to forecast UTLS turbulence for aviation purposes. Specifically, we target a more flexible ML-based GTG-like algorithm that (i) overcomes some of the weaknesses of the current GTG optimization method, (ii) provides an alternative to the EDR mapping approach, (iii) results in an increased predictive skill relative to the additional complexity required, and (iv) is sufficiently computationally efficient to be incorporated into a real-time GTG-like prediction system. A database of 2.42 million model–observation pairs was created, as the result of 1089 three-km HRRR-based GTG 6-h forecasts, randomly split into training (75%) and testing (25%) datasets. After performing sensitivity analyses to two of the hyperparameters of the random forest (RF) ensemble technique, namely, the number of

trees n_t and the tree depth t_d , it was found that an algorithm with $n_t = 100$ trees of maximum $t_d = 30$ layers of depth provides sufficient complexity for this ML method for the specific problem considered herein. This baseline RF-based algorithm is found to significantly reduce forecast errors especially for $\text{EDR} < 0.1 \text{ m}^{2/3} \text{ s}^{-1}$ by increasing the probability of detection and in turn reducing the number of false alarms. This EDR range is representative of null and light turbulence for a typical medium-sized aircraft, where the HGTG algorithm tends to overpredict turbulence intensity. Accurately predicting calm conditions is also a very important aspect of a turbulence forecasting algorithm. This allows airlines to avoid overusing the seat-belt-on sign, which in turn improves customer perception of reliability and permits the cabin crew to normally operate their aircraft cart services.

We also explore a reduction of the complexity of the baseline RF model by employing gradient-boosted regression trees and two feature reduction methods based on relative importance identification by impurity (RFri) and by permutation (RFrp). In particular, we find that the number of features can be considerably reduced, from 114 in the baseline model to 32 in reduced models where 60% of the total impurity importance (RFri60) and 89% of the total permutation importance (RFrp89) are retained while not noticeably impacting the model's skill. In fact, the performance of the reduced models for observed $\text{EDR} < 0.15 \text{ m}^{2/3} \text{ s}^{-1}$ is slightly improved with respect to the baseline RF model, an aspect attributed to a reduction in overfitting and hence a more generalizable model. The GBRT model produces a similar skill to that of the RF baseline model, in addition to speeding up the execution of the algorithm by a factor of ≈ 3.75 . The reduced feature RF-based models do not directly impact the speed of execution of the ML algorithm (same number of trees and layers); however, they do indirectly decrease the global execution time through an effective reduction in the number of features that are input to the RF algorithm.

The ML-based models overall exhibit augmented average TSS for light–moderate–severe turbulence, in turn leading to higher AUC values of 0.87 for the RFri60 model versus 0.74 for the HGTG algorithm, indicative of a higher skill in discriminating the EDR forecast among these categories. Another advantage of these ML-based techniques is that they directly use the turbulence diagnostics, therefore eliminating the need for performing a calibration to fit to a climatological EDR distribution or mapping to EDR space, as required by the current GTG method. These characteristics, as already mentioned, considerably contribute to streamlining the generation of NWP and grid-spacing-specific GTG forecast products. Feature-relative-importance analyses were used to help interpret the results of the ML model. These results imply that the characteristics of the environment surrounding the tropopause are typically the most important contributing factor to UTLS turbulence events. Variables related to moist convection are found to also be important features, with a lower contribution compared to the tropopause-related mechanisms, likely related to a bias in the available data originated from the avoidance of moist convection detectable by onboard radars when possible. Also, resolved small-scale dynamics are of relevance, supporting the notion that high-resolution NWP models are

successfully capturing turbulence drivers. These insights obtained from the ML model are useful to guide future efforts in the development of enhanced turbulence diagnostics.

We have demonstrated that RF models can be employed to improve GTG's predictive skill, in particular to reduce the frequency of false alarms in null or very light turbulence conditions, and also simplifies the current GTG framework. Based on these promising findings, we plan to incorporate this type of RF algorithm within the GTG framework for its operational use, including a specific product for mountain wave turbulence and the development of a probabilistic forecast product initially derived from the predictions by the individual trees of the RF model. In addition, we will explore the extension of the current ML-based algorithm to forecast low-level (Muñoz-Esparza and Sharman 2018) and convectively induced turbulence. For the latter, we plan to employ observations of in-cloud EDR from the NCAR/NEXRAD Turbulence Detection Algorithm (NTDA; Williams and Meymaris 2016). This additional observational data source provides a three-dimensional in-cloud mosaic of EDR produced every 5 min over the contiguous United States, thereby complementing the sparser within-cloud in situ and PIREPs datasets.

One aspect worth keeping in mind when developing these types of algorithms is the particular nature of turbulence forecasting using coarse NWP model data that we are dealing with. On one hand, we are attempting to forecast turbulence at the subgrid, which would minimally require the proper combination of turbulence diagnostics for each particular weather situation, and that is inherently limited by the ability of performing such complex downscaling operation. On the other hand, we are basing our model development/training on NWP model data, with its own inherent forecast errors. These two combined do in fact result in a very challenging problem, since we are targeting the development of an accurate model yet trained with imperfect forecasts, while at the same time seeking to infer subgrid-scale effects for which a known universal relationship that is applicable to all particular weather situations does not exist. Also, the observations that our turbulence prediction algorithms are based upon, that is, automated in situ EDR and PIREPs are only available at the times, locations, and altitudes where these commercial aircraft operate, often attempting to avoid regions of turbulence activity, which further limits the coverage of the reference dataset. Therefore, there are combined intricacies that cannot simply be untangled and that affect the process of developing these turbulence forecast algorithms as a whole. Still, GTG and similar algorithms have proven to be a useful source of information for both strategic and tactical turbulence avoidance (Sharman and Lane 2016), therefore confirming the need for sustained research in aviation turbulence forecasting, effectively combining innovative algorithms, high-resolution models, and comprehensive observational networks.

Acknowledgments. This research is in response in part to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA. The authors are grateful to Teddie Keller for her help in selecting the cases presented in section 6 and to Matthias

Steiner, Jeremy Sauer, and Branko Kosović for providing feedback on an early version of the paper. All of the GTG simulations and ML model training/development were performed using high-performance computing support from Cheyenne (<https://doi.org/10.5065/D6RX99HX>) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.

REFERENCES

- Abernethy, J., R. Sharman, and E. Bradley, 2008: An artificial intelligence approach to operational aviation turbulence forecasting. *Third Int. Conf. on Research in Air Transportation*, Fairfax, VA, FAA and EUROCONTROL, 429–435.
- Ahijevych, D., J. O. Pinto, J. K. Williams, and M. Steiner, 2016: Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. *Wea. Forecasting*, **31**, 581–599, <https://doi.org/10.1175/WAF-D-15-0113.1>.
- Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, <https://doi.org/10.1175/MWR-D-15-0242.1>.
- Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, <https://doi.org/10.1023/A:1010933404324>.
- Brown, R., 1973: New indices to locate clear-air turbulence. *Meteor. Mag.*, **102**, 347–361.
- Burke, A., N. Snook, D. J. Gagne II, S. McCorkle, and A. McGovern, 2020: Calibration of machine learning-based probabilistic hail predictions for operational forecasting. *Wea. Forecasting*, **35**, 149–168, <https://doi.org/10.1175/WAF-D-19-0105.1>.
- Chandler, C. L., 1987: Turbulence forecasting. *Atmospheric Turbulence Relative to Aviation, Missile, and Space Programs*, D. W. Camp and W. Frost, Eds., NASA Conference Publ. 2468, 137–154.
- Danielsen, E. F., 1968: Stratospheric-tropospheric exchange based on radioactivity, ozone and potential vorticity. *J. Atmos. Sci.*, **25**, 502–518, [https://doi.org/10.1175/1520-0469\(1968\)025<0502:STEBOR>2.0.CO;2](https://doi.org/10.1175/1520-0469(1968)025<0502:STEBOR>2.0.CO;2).
- Ellrod, G. P., and D. I. Knapp, 1992: An objective clear-air turbulence forecasting technique: Verification and operational use. *Wea. Forecasting*, **7**, 150–165, [https://doi.org/10.1175/1520-0434\(1992\)007<0150:AOCATF>2.0.CO;2](https://doi.org/10.1175/1520-0434(1992)007<0150:AOCATF>2.0.CO;2).
- Endlich, R. M., 1964: The mesoscale structure of some regions of clear-air turbulence. *J. Appl. Meteor.*, **3**, 261–276, [https://doi.org/10.1175/1520-0450\(1964\)003<0261:TMSOSR>2.0.CO;2](https://doi.org/10.1175/1520-0450(1964)003<0261:TMSOSR>2.0.CO;2).
- Frehlich, R., and R. Sharman, 2004: Estimates of turbulence from numerical weather prediction model output with applications to turbulence diagnosis and data assimilation. *Mon. Wea. Rev.*, **132**, 2308–2324, [https://doi.org/10.1175/1520-0493\(2004\)132<2308:EOTFNW>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<2308:EOTFNW>2.0.CO;2).
- Friedman, J. H., 2001: Greedy function approximation: A gradient boosting machine. *Ann. Stat.*, **29**, 1189–1232, <https://doi.org/10.1214/aos/1013203451>.
- Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017a: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, <https://doi.org/10.1175/WAF-D-17-0010.1>.
- , —, —, and J. K. Williams, 2017b: Evaluation of statistical learning configurations for gridded solar irradiance forecasting. *Sol. Energy*, **150**, 383–393, <https://doi.org/10.1016/j.solener.2017.04.031>.

- , T. C. McCandless, T. Brummet, B. Kosović, and S. E. Haupt, 2019: Surface layer flux machine learning parameterizations. *18th Conf. on Artificial and Computational Intelligence and Its Applications to the Environmental Sciences*, Phoenix, AZ, Amer. Meteor. Soc., 5B.1, <https://ams.confex.com/ams/2019Annual/webprogram/Paper352862.html>.
- Gill, P. G., 2014: Objective verification of World Area Forecast Centre clear air turbulence forecasts. *Meteor. Appl.*, **21**, 3–11, <https://doi.org/10.1002/met.1288>.
- , 2016: Aviation turbulence forecast verification. *Aviation Turbulence: Processes, Detection, Prediction*, R. Sharman and T. Lane, Eds., Springer, 261–283.
- Herman, G. R., and R. S. Schumacher, 2018: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, <https://doi.org/10.1175/MWR-D-17-0250.1>.
- Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Mon. Wea. Rev.*, **148**, 2135–2161, <https://doi.org/10.1175/MWR-D-19-0344.1>.
- ICAO, 2001: *Meteorological Service for International Air Navigation. Annex 3, Convention on International Civil Aviation*. 14th ed., ICAO, 128 pp.
- Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 274 pp.
- Kim, S.-H., H.-Y. Chun, J.-H. Kim, R. D. Sharman, and M. Strahan, 2020: Retrieval of eddy dissipation rate from derived equivalent vertical gust included in Aircraft Meteorological Data Relay (AMDAR). *Atmos. Meas. Tech.*, **13**, 1373–1385, <https://doi.org/10.5194/amt-13-1373-2020>.
- Koch, S. E., and Coauthors, 2005: Turbulence and gravity waves within an upper-level front. *J. Atmos. Sci.*, **62**, 3885–3908, <https://doi.org/10.1175/JAS3574.1>.
- Kosović, B., and Coauthors, 2020: A comprehensive wind power forecasting system integrating artificial intelligence and numerical weather prediction. *Energies*, **13**, 1372, <https://doi.org/10.3390/en13061372>.
- Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine learning for real-time prediction of damaging straight-line convective wind. *Wea. Forecasting*, **32**, 2175–2193, <https://doi.org/10.1175/WAF-D-17-0038.1>.
- Lester, P., 1993: *Turbulence: A New Perspective for Pilots*. Jeppesen Sanderson, 280 pp.
- Lindborg, E., 1999: Can the atmospheric kinetic energy spectrum be explained by two-dimensional turbulence? *J. Fluid Mech.*, **388**, 259–288, <https://doi.org/10.1017/S0022112099004851>.
- Louppe, G., L. Wehenkel, A. Sutera, and P. Geurts, 2013: Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems 26*, C. Burges, Ed., Curran, 431–439.
- Marroquin, A., 1998: An advanced algorithm to diagnose atmospheric turbulence using numerical model output. *16th Conf. on Weather Analysis and Forecasting*, Phoenix, AZ, Amer. Meteor. Soc., 79–81.
- McCann, D. W., 2001: Gravity waves, unbalanced flow, and aircraft clear air turbulence. *Natl. Wea. Dig.*, **25**, 3–14.
- McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, <https://doi.org/10.1175/BAMS-D-16-0123.1>.
- , C. D. Karstens, T. Smith, and R. Lagerquist, 2019a: Quasi-operational testing of real-time storm-longevity prediction via machine learning. *Wea. Forecasting*, **34**, 1437–1451, <https://doi.org/10.1175/WAF-D-18-0141.1>.
- , R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019b: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, <https://doi.org/10.1175/BAMS-D-18-0195.1>.
- Muñoz-Esparza, D., and R. Sharman, 2018: An improved algorithm for low-level turbulence forecasting. *J. Appl. Meteor. Climatol.*, **57**, 1249–1263, <https://doi.org/10.1175/JAMC-D-17-0337.1>.
- Partl, W., 1962: Clear air turbulence at the tropopause levels. *Navigation*, **9**, 288–295, <https://doi.org/10.1002/j.2161-4296.1962.tb02532.x>.
- Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
- Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, <https://doi.org/10.1175/MWR-D-18-0187.1>.
- Reap, R., 1996: Probability forecasts of clear-air-turbulence for the contiguous U.S. National Weather Service Office of Meteorology Tech. Procedures Bull. 430, 15 pp.
- Roach, W., 1970: On the influence of synoptic development on the production of high level turbulence. *Quart. J. Roy. Meteor. Soc.*, **96**, 413–429, <https://doi.org/10.1002/qj.49709640906>.
- Scikit-Learn-Developers, 2018: Scikit-learn user guide. Release 0.19.2, 214–215.
- Shapiro, M., 1978: Further evidence of the mesoscale and turbulent structure of upper level jet stream-frontal zone systems. *Mon. Wea. Rev.*, **106**, 1100–1111, [https://doi.org/10.1175/1520-0493\(1978\)106<1100:FEOTMA>2.0.CO;2](https://doi.org/10.1175/1520-0493(1978)106<1100:FEOTMA>2.0.CO;2).
- , 1980: Turbulent mixing within tropopause folds as a mechanism for the exchange of chemical constituents between the stratosphere and troposphere. *J. Atmos. Sci.*, **37**, 994–1004, [https://doi.org/10.1175/1520-0469\(1980\)037<0994:TMWTF>2.0.CO;2](https://doi.org/10.1175/1520-0469(1980)037<0994:TMWTF>2.0.CO;2).
- Sharman, R., and T. Lane, 2016: *Aviation Turbulence: Processes, Detection, Prediction*. Springer, 523 pp.
- , and J. Pearson, 2017: Prediction of energy dissipation rates for aviation turbulence. Part I: Forecasting nonconvective turbulence. *J. Appl. Meteor. Climatol.*, **56**, 317–337, <https://doi.org/10.1175/JAMC-D-16-0205.1>.
- , and S. Trier, 2019: Influences of gravity waves on convectively induced turbulence (CIT): A review. *Pure Appl. Geophys.*, **176**, 1923–1958, <https://doi.org/10.1007/s00024-018-1849-2>.
- , C. Tebaldi, G. Wiener, and J. Wolff, 2006: An integrated approach to mid- and upper-level turbulence forecasting. *Wea. Forecasting*, **21**, 268–287, <https://doi.org/10.1175/WAF924.1>.
- , S. Trier, T. Lane, and J. Doyle, 2012: Sources and dynamics of turbulence in the upper troposphere and lower stratosphere: A review. *Geophys. Res. Lett.*, **39**, L12803, <https://doi.org/10.1029/2012GL051996>.
- , L. Cornman, G. Meymaris, J. Pearson, and T. Farrar, 2014: Description and derived climatologies of automated in situ eddy-dissipation-rate reports of atmospheric turbulence. *J. Appl. Meteor. Climatol.*, **53**, 1416–1432, <https://doi.org/10.1175/JAMC-D-13-0329.1>.
- Skamarock, W. C., and J. B. Klemp, 2008: A time-split non-hydrostatic atmospheric model for weather research and forecasting applications. *J. Comput. Phys.*, **227**, 3465–3485, <https://doi.org/10.1016/j.jcp.2007.01.037>.
- Smith, T. L., S. G. Benjamin, J. M. Brown, S. Weygandt, T. Smirnova, and B. Schwartz, 2008: Convection forecasts

- from the hourly updated, 3-km High Resolution Rapid Refresh (HRRR) model. *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 11.1, <https://ams.confex.com/ams/pdfpapers/142055.pdf>.
- Trier, S. B., R. D. Sharman, and D. Muñoz-Esparza, 2020: Environment and mechanisms of severe turbulence in a mid-latitude cyclone. *J. Atmos. Sci.*, **77**, 3869–3889, <https://doi.org/10.1175/JAS-D-20-0095.1>.
- Williams, J. K., 2014: Using random forests to diagnose aviation turbulence. *Mach. Learn.*, **95**, 51–70, <https://doi.org/10.1007/s10994-013-5346-7>.
- , and G. Meymaris, 2016: Remote turbulence detection using ground-based Doppler weather radar. *Aviation Turbulence: Processes, Detection, Prediction*, R. Sharman and T. Lane, Eds., Springer, 149–177.
- , D. Ahijevych, S. Dettling, and M. Steiner, 2008: Combining observations and model data for short-term storm forecasting. *Proc. SPIE*, **7088**, 708805, <https://doi.org/10.1117/12.795737>.
- Wolff, J., and R. Sharman, 2008: Climatology of upper-level turbulence over the contiguous United States. *J. Appl. Meteor. Climatol.*, **47**, 2198–2214, <https://doi.org/10.1175/2008JAMC1799.1>.