Multi-model ensemble predictions of aviation turbulence

Article

Accepted Version

Multi-Model Ensemble Predictions of Aviation Turbulence

Multi-model predictions of Turbulence

Luke N. Storer*[1], Philip G. Gill[2] and Paul D. Williams[1]

1. Department of Meteorology, University of Reading, Reading, UK
2. Met Office, Exeter, UK

***Corresponding author:** *luke.storer@pgr.reading.ac.uk*

Abstract

Turbulence remains one of the leading causes of aviation incidents. Climate change is predicted to increase the occurrence of Clear-Air Turbulence (CAT), and therefore forecasting turbulence will become more important in the future. Currently the two World Area Forecast Centres (WAFCs) use deterministic numerical weather prediction models to predict clear-air turbulence operationally, it has been shown that ensemble forecasts improve the forecast skill of traditional meteorological variables. This study applies multi-model ensemble forecasting to aviation turbulence for the first time. It is shown in a 12-month global trial from May 2016 to April 2017, that combining two different ensembles yields a similar forecast skill to a single model ensemble, and yields an improvement in forecast value at low cost/loss ratios. This finding is consistent with previous work showing that the use of ensembles in turbulence forecasting is beneficial. Using a multi-model approach is an effective way to improve the forecast skill and provide pilots and flight planners with more information about the forecast confidence, allowing them to make a more informed decision about what action needs to be taken, such as diverting around the turbulence or requiring passengers and flight attendants to be seatbelted. The multi-model ensemble approach is intended to be made operational by both WAFCs in the near future and this study lays the foundations to make this possible.

Keywords: Turbulence, Aviation, Multi-model, Ensemble, Forecast

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/met.1772

1. Introduction

Aviation turbulence is experienced by most people who fly. It is a major hazard, with tens of millions of dollars paid out annually by airline companies to passengers and crew who are injured, and over 7,000 person-days of lost time for cabin crew related injuries [Sharman and Lane, 2016]. Therefore forecasting turbulence is vital in maintaining the safety of passengers and crew. Turbulence affecting aviation comes in different forms including Clear-Air Turbulence (CAT), which is particularly hazardous because it cannot be remotely detected by pilots. CAT is defined as high-altitude in-flight bumps in airspace devoid of significant cloudiness and away from thunderstorm activity [Chambers, 1955]. Turbulence can be formed by Mountain Wave Turbulence (MWT) [Lilly, 1978], Convectively Induced Turbulence (CIT) [Uccellini and Koch, 1987; Koch and Dorian, 1988] and shear-induced turbulence [Endlich, 1964; Atlas *et al.*, 1970]. Our ability to forecast turbulence has improved recently, because of advances in our mechanistic understanding [Williams *et al.*, 2003, 2005, 2008; Knox *et al.*, 2008; McCann *et al.*, 2012] and new measurement techniques [Marlton *et al.*, 2015]. An extensive overview of aviation turbulence can be found in Sharman and Lane [2016].

There are two World Area Forecast Centres (WAFC) --- London (Met Office) and Washington (NOAA) --- that create the turbulence forecasts used operationally by pilots and flight planners around the world. Currently the two centres produce turbulence forecasts by creating a turbulence product with a horizontal resolution of 1.25° from a single output deterministic model. Turbulence cannot be explicitly simulated by numerical weather prediction models, because the turbulence scales impacting aviation are between 100 m and 1 km, which is smaller than a grid box of an operational global forecast system [Sharman *et al.*,

2006]. The WAFCs therefore use a diagnostic indicator to predict areas of the atmosphere likely to contain turbulence. Both WAFCs use the Ellrod and Knapp [1992] Turbulence Index 1 (Ellrod TI1) [ICAO, 2012], which predicts shear-induced turbulence that is dominated by strong deformation regions associated with the jet stream. It is important to note that the Ellrod TI1 is not able to capture all shear turbulence generation mechanisms. However, the jet stream is a current of fast-flowing air that is constantly evolving and difficult to predict. Therefore, a deterministic model arguably does not fully capture the uncertainty of the parameters of the jet stream and the location of turbulence [Gill and Buchanan, 2014]. Also the Ellrod TI1 is unable to detect convective events or mountain wave turbulence unless in areas of strong wind shear. Gill [2014] demonstrates this by showing the skill of the TI1 is reduced in the Tropics where shear turbulence is less important, and convection is the main cause of turbulence. Ellrod and Knox [2010] developed an improvement for the Ellrod TI1 turbulence index by including a divergence trend. This has resulted in the Ellrod3 turbulence diagnostic which Sharman and Pearson [2017] showed to have a higher area under the curve than other turbulence diagnostics. We choose the Ellrod TI1 in this study to be consistent with the current operational forecast system, however, this shows there is potential to further increase forecast skill by using this updated turbulence index.

To help include jet stream uncertainty in the forecasts, Gill and Buchanan [2014] and Buchanan [2016] showed that using an ensemble of simulations improved the forecast skill. An ensemble forecast is a collection of many simulations of the same event with each outcome equally likely. By combining all the possible results, a probability field showing the likelihood of exceeding a turbulence threshold and therefore experiencing turbulence can be created. Having a probabilistic forecast can help the pilots, flight planners and Air Traffic Control (ATC) manage their response accordingly. An example of this would be if 1 out of 10 ensemble members predict turbulence (i.e. there is a 10 % probability of turbulence), in

which case pilots might continue their route because the chances are still small. If, on the other hand, all 10 models predict that the threshold will be exceeded, then a pilot may divert around that region (expensive), change flight level (less expensive), or put the seat belt sign on (free) to avoid injury to passengers and crew. Choosing the appropriate action can reduce injuries and save costs. For example, if the turbulence is expected to be light or moderate or there is a low probability of the turbulence, then putting the seat belt sign on is a cost-free response but can impact passenger comfort. If the turbulence predicted is severe or there is a high probability, then the pilot might choose to change the flight level, which might cost money in terms of fuel usage but this cost would be less than a full diversion. If the turbulence predicted is on multiple flight levels, then a full diversion might be appropriate, which would be more expensive by increasing flight time and fuel usage, but this would be cheaper than damaging the aircraft or injuring passengers and crew.

Having a probabilistic forecast with more information allows pilots and flight planners to choose the appropriate action to reduce the costs of preventative action whilst maintaining the safety of passengers and crew. If more ensemble members are used, the ensemble spread is larger and the understanding of the certainty of the forecast is improved. This approach therefore provides more information to pilots and flight planners about where turbulence is likely to be and therefore which regions they should avoid. However, increasing the forecast spread could capture more turbulence events, but also could increase the number of false alarms, and this trade-off is one that needs to be managed to maximise forecast skill. This study further expands on the use of an ensemble forecast, and follows other areas of meteorology such as the TIGGE project [Swinbank *et al.*, 2016] which looks at combining ensembles from different centres around the world. A particular research area using multi-model ensembles is tropical cyclone forecasting [Krishnamurti *et al.*, 2000; Vitart, 2006; Titley and Stretton, 2016]. All of these studies show that using multi-model ensembles

improves the overall skill of the forecasts, and they therefore show a useful application that we will investigate for turbulence in this study. By using at least two ensembles, not only is the spread increased by increasing the number of forecasts, but also a different numerical model is used, assimilating a differing set of observations, that will have different strengths and weaknesses. These strengths and weaknesses can come from how the ensembles are perturbed. An example is the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS), which starts each model run with the same initial conditions but adds dynamically defined perturbations to create the model spread [Molteni *et al.*, 1996]. In contrast, the Met Office Global and Regional Ensemble Prediction System (MOGREPS-G) uses different initial conditions and model perturbations to provide the ensemble spread [Bowler *et al.*, 2008]. The initial conditions are perturbed using the ensemble transform Kalman filter [Bishop *et al.*, 2001] and the model perturbations are driven by two stochastic physics schemes. These two schemes are the random parameter scheme and the stochastic convective vorticity scheme. Bowler *et al.* [2008] showed as an example that the screen temperature Brier Skill Score (BSS) was higher for ECMWF-EPS compared to MOGREPS-G, but that for wind speed the MOGREPS-G ensemble was more skilful than the ECMWF-EPS. This shows each ensemble has its own strengths and weaknesses that we hope will increase the forecast spread and therefore increase the forecast skill. This study is the first time multi-model ensemble forecasting has been applied to turbulence. Both WAFCs plan to use a multi-model ensemble in the near future, and therefore this study lays the foundations to make this possible.

Turbulence forecasting will become more important in the future, because climate change is predicted to increase the frequency of clear-air turbulence globally [Storer *et al.*, 2017]. The turbulence increases arise because of changes to the jet streams, which are also predicted to modify flight routes and journey times [Williams, 2016]. Storer *et al.* [2017] showed that

flights all around the world will have an increase in turbulence for all strength categories from light to severe. Therefore, improved turbulence forecasts will be a vital tool to limit the increase in injuries to passengers and crew as well as aircraft damage arising from the increase in turbulence events. Not only will this improve passenger comfort and safety, it will also reduce the loss of money paid out for compensation.

This paper is set out as follows. Section 2 will introduce the observational data, section 3 will explain the forecasting system, section 4 will discuss the verification method, section 5 will present the results, and section 6 will summarise the results and outline future work.

2. Observations

In order to verify the forecasts, we need to find a 'truth' data set. Previous work used PIlot REPortS (PIREPS) [Tebaldi *et al.*, 2002; Kim and Chun, 2011], but these can be unreliable [Kane *et al.*, 1998; Schwartz, 1996; Sharman *et al.*, 2014]. PIREPS are subjective and are also aircraft dependent, so a smaller aircraft will experience more severe turbulence than a larger aircraft in the same volume of turbulent air. PIREPS also have poor spatial reliability as they tend to be located in turbulence, so null turbulence events are rarely recorded as there is no specified frequency [Kane *et al.*, 1998]. The location and time of PIREPS may also not be correct as they are manually reported after the event and for more severe events where action is required this may be some time later. To avoid these problems, this study will use aircraft data recorded on a fleet of Boeing 747 and 777 aircraft. This data has been used in other meteorological studies [Tenenbaum, 1991; Gill, 2014]. High-resolution automated aircraft data, available at 4 second intervals, giving us over 76, 000, 000 data points, are used to calculate an aircraft-independent turbulence measure known as the Derived Equivalent Vertical Gust (DEVG) which is defined as:

$$\text{DEVG} = \frac{Am|\Delta n|}{V} \qquad (1)$$

…where Δn is the peak modulus value of deviation of aircraft acceleration from $1g$ in units of g, m is the total mass of the aircraft (metric tonnes), V is the calibrated airspeed at the time of the observation (knots), and A is an aircraft specific parameter which varies with flight conditions and is defined as:

$$A = \bar{A} + c_4(\bar{A} - c_5)\left(\frac{m}{\bar{m}-1}\right) \quad (2)$$

$$\bar{A} = c_1 + \frac{c_2}{c_3 + H(kft)} \quad (3)$$

…where H is the altitude (thousands of feet), $\bar{m}$ is the reference mass of the aircraft (metric tonnes), and parameters c_1 to c_5 depend on the aircraft's flight profile as outlined in Truscott [2000].

DEVG is one of the World Meteorological Organization (WMO) recommended turbulence indicators and has a typical uncertainty of around 3-4 % [WMO, 2003]. DEVG is aircraft independent so values from all aircraft can be combined to create an observational database. Table 1 compares DEVG to EDR which is another aircraft independent measure that can both be used in turbulence verification [Storer et al. 2018]. There are limitations to using this data set, because aircraft manoeuvers and active control techniques can enhance or dampen vertical accelerations of aircraft leading to over- or under-representation of the vertical gusts [WMO, 2003]. One of the other main issues with this data set is the typical spatial coverage. Figure 1 is a plot of aircraft data for May 2016 and shows the spatial coverage of our observations. It has very good coverage over the North Atlantic and Europe, but poorer coverage over Asia and the Pacific. Despite the uneven spatial coverage, this data set is still the best available source of truth data for verification, which is why we have chosen to use it here. The Ellrod TI1 turbulence predictor used in this study only forecasts shear induced turbulence and is not able to predict convective turbulence. This study will therefore use a satellite-based convective product to filter out convective turbulence events [Francis and

Batstone, 2013]. By only looking at the non-convective events we should have a better representation of the forecast skill by removing events that will not be forecast in this study.

3. Forecast data

This project uses an entire year of global ensemble data between May 2016 and April 2017 from two forecast centres: MOGREPS-G [Bowler *et al.*, 2008] and the ECMWF EPS [Molteni *et al.*, 1996]. The forecast data is available with 3-hourly intervals and at the time of this study the MOGREPS-G ensemble had 12 members with forecasts every 6 hours, with 33 km resolution and 70 vertical levels, 10 of which are between 150 and 350 hPa. The ECMWF forecast had 51 ensemble members with 18 km resolution and 91 vertical levels, 14 of which are between 150 and 350 hPa. To use the ECMWF EPS system operationally we would have to extend the forecast range by approximately 12 hours due to a delay in accessing the forecast data. This is important to note because it means the results of the ECMWF EPS will be theoretical and in practice the skill would be reduced as we have to use longer lead times. This is shown by the WAFC verification website demonstrating how the forecast skill is reduced with forecast lead time [Met Office, 2018]. We have not used the longer lead times in this study but understanding the impact this will have on the forecast skill would be an interesting area of further study. Also for both ensembles this study only had access to the 0000 UTC model run between 01 May 2016 and 07 August 2016, which means for that period we forecast only half of the day. Between 07 August 2016 and 30 April 2017 we had both the 0000 UTC and 12000 UTC model run. The forecast lead time used throughout is T+24, T+27, T+30, T+33 hours.

This study focuses on non-convective turbulence in this paper, and uses the Ellrod TI1 predictor which is defined as:

$$\mathrm{TI1} = DEF \times VWS = \left\{\left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right)^2 + \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)^2\right\}^{\frac{1}{2}} \times \left\{\left(\frac{\partial u}{\partial z}\right)^2 + \left(\frac{\partial v}{\partial z}\right)^2\right\}^{\frac{1}{2}} \quad (4)$$

…where u is the horizontal wind velocity in the East-West direction; v is the horizontal wind velocity in the North-South direction; x is distance in the East-West direction; y is distance in the North-South direction; and z is distance in the vertical. This is the same index used in

previous studies and is the turbulence diagnostic currently used by the WAFCs. The Ellrod TI1 combines deformation and vertical wind-shear as shown in Equation (4) and is a well-established shear turbulence diagnostic. Previous research has shown it predicts up to 65 % of CAT events [Sharman *et al.*, 2006], although it is typically only useful in the mid-to-high latitudes. Also the Ellrod TI1 was not developed to predict CIT or MWT, which are both prominent turbulence sources.

To create a probability forecast, we use the ensemble data from both WAFCs and set thresholds based on literature [ICAO, 2012]. The turbulence threshold used in this project for the Ellrod TI1 is $3 \times 10^{-7} s^{-2}$ which is equivalent to Moderate or Greater (MoG) turbulence. However we also calculate additional thresholds that range from Light or Greater to Severe or Greater turbulence. Using multiple thresholds to predict different turbulence strength categories is similar to the approach used by Williams [2017] and Storer et al. [2017]. Ellrod and Knapp [1992] also discuss the use of higher thresholds for moderate and severe turbulence with the actual values used being model specific. We use these additional thresholds to optimise the turbulence forecast as a lower threshold will capture more MoG turbulence events, but also more false alarms, and a higher threshold will forecast fewer MoG events but also fewer false alarms. The additional thresholds we used are $8 \times 10^{-8} s^{-2}$, $1 \times 10^{-7} s^{-2}$, $5 \times 10^{-7} s^{-2}$, $8 \times 10^{-7} s^{-2}$, $1.1 \times 10^{-6} s^{-2}$, and $2 \times 10^{-6} s^{-2}$. Above these thresholds, it is classed as an area of the atmosphere containing turbulence. By combining all the ensemble results it gives the probability for a grid point containing MoG turbulence. The more ensemble forecasts that predict the occurrence of turbulence, the higher the probability forecast will be. It is then possible to combine both the ensemble forecasts, which can be done in two ways. The first is a standard equally weighted multi-model super ensemble, and the other is a weighted multi-model ensemble. This study uses the simple super ensemble by combining the two equally weighted ensembles. This was created by first calculating the probability

field of turbulence for both the single-model ensembles based on exceeding a threshold of $3 \times 10^{-7} s^{-2}$. Then the mean of the probability for both ensembles created the multi-model ensemble. This therefore means that although the ECMWF-EPS ensemble has more ensemble members, it does not have any more weight in the multi-model ensemble.

4. Verification method

The verification method used in this study is outlined in Gill [2014, 2016]. This study processes aircraft observations into 10-minute segments, which equates to approximately 100 km of flight. By analysing the DEVG values in each segment, if the maximum value exceeds a given threshold it is classed as a turbulence event. The aircraft data are constrained to ±1.5 hours of the forecast time to ensure the aircraft observations reflect the forecast. Therefore each 10 minute segment (observation) within the 3 hour time window is used for verification as beyond this the observations are not valid for the forecast. It is then possible to compare the turbulence observations to the forecast and a 2 × 2 contingency table can be set up as shown in Table 2. One of the best ways to visualize these results is to use a Relative Operating Characteristic (ROC) plot [Jolliffe and Stephenson, 2012; Gill, 2016], which plots the hit rate against the false alarm rate which is defined as:

$$\text{Hit Rate (H)} = \frac{A}{A+C} \quad (5)$$

$$\text{False Alarm Rate (F)} = \frac{B}{B+D} \quad (6)$$

…where A is a hit; B is a false alarm; C is a miss; and D is a correct rejection. To create the ROC curve, thresholds are applied to the probabilities which then create binary yes/no forecasts with corresponding 2x2 contingency tables yielding the hit rate and false alarm rate which are then plotted together. This produces a curve where the larger the area under the curve, the more skilful the forecast is at discriminating between events and non-events.

The reliability of the forecasts can be assessed visually by using a reliability diagram [Jolliffe and Stephenson, 2012; Gill, 2016], where each probability is binned and the frequency of the event is calculated. The forecast probability should equal the observed frequency. For example, if the probability is 30 % then turbulence in that region should be observed 30 % of the time. Therefore a perfect forecast would result in a straight line, however in practice this is not the case and forecast probabilities tend to over-forecast the turbulence (below the line), or under-predict turbulence (above the line). Understanding these biases allow us to implement a linear calibration which should bring the forecast probability more in-line with the observed frequency. Calibrating the forecast will not compromise the forecast skill, since ROC area discriminatory skill and reliability are independent [Gill, 2016].

A more practical analysis of the results for stakeholders would be to assess the relative economic value (V) of the forecast which is defined as:

$$V = \frac{\min(\alpha,\bar{o}) - F\alpha(1-\bar{o}) + H\bar{o}(1-\alpha) - \bar{o}}{\min(\alpha,\bar{o}) - \bar{o}\alpha} \qquad (7)$$

…where α = Cost/Loss, $\bar{o}$ is the fraction of occasions where the event occurred, F is the false alarm rate, H is the hit rate [Richardson, 2000; Jolliffe and Stephenson, 2012]. This assigns a cost and loss for the elements in a contingency table (Table 3) where different outcomes depend on whether action was taken and if the event occurred or not. For a given model, the hit rate (H), false alarm rate (F) and fraction of occasions the event occurred $\bar{o}$ can be calculated using a 2x2 contingency table, and therefore varying the cost/ loss ratio gives a different value which can be plotted. The more skilful the model, the higher the maximum value will be (but the actual value will depend on the cost/ loss ratio of the user), and if the value is higher for all cost/ loss ratios then that model will be the most useful for any consumer (as the cost/ loss ratio may vary depending on the consumer) and this is known as sufficiency [Ehrendorfer and Murphy, 1988]. Gill and Buchanan [2014] and Buchanan [2016] showed that probability turbulence forecasts have greater value than deterministic

forecasts, so this project will aim to show by combining ensembles that we can further increase the value.

5. Results

Throughout this analysis we focus on shear turbulence, however MWT and CIT will be present in the observational truth data. To identify the source of turbulence, this study plots the aircraft data over a plot of orography, convection, and shear turbulence. The orography plot uses a surface map that indicates terrain height and therefore mountain ranges. The orography map shows the height of the terrain, and any event that occurs near high terrain it could be caused by MWT. Whereas the CIT plot uses a satellite product that indicates areas of deep convection [Francis and Batstone, 2013]. The satellite product identifies regions of overshooting tops which indicate the regions of the strongest updraft, above the smooth anvil of a typical thunderstorm. To identify these regions they use two methods, the first method is the water vapour-infrared window brightness temperature difference method [Schmetz et al., 1997]. The second is the infrared window texture method [Bedka et al. 2010]. By using the infrared channel, it can be used in both the day and night which is important for aviation. We did not have full global coverage for the satellite product, and therefore only CIT events within that spatial coverage could be removed. For the shear turbulence we plot both ensemble probability fields so it will show if shear turbulence is a likely cause and if both ensemble products predict turbulence. The plots have aircraft data ±1.5 hours which will help identify the likely source of turbulence.

Figure 2 is a plot of a shear turbulence case study that was forecast by both the MOGREPS-G and ECMWF-EPS ensembles. The plot clearly shows the MoG turbulence event was over the North Atlantic so MWT is not a factor, and the satellite product shows there is no deep convection in the area, although there is some convection much further south. This shows the

turbulence event was well forecast by the ensemble products. There is also a light turbulence event further south, and this is not forecast by either ensemble. This is to be expected as the threshold used for this figure is typical for MoG turbulence, and therefore not expected for this event. There is some convection just to the south, however it seems too far away to cause this light turbulence event.

Looking at Figure 3 there is no CIT in the area however this could be MWT as this event is over some smaller mountains, or shear turbulence as it is forecast by the MOGREPS-G ensemble. This is a case that shows the benefit of using the multi-model ensemble approach. If we only had the ECMWF-EPS ensemble then we would not forecast this event and as a result people could be injured. But since we have both the ensembles, we have forecasted the event and therefore preventative action could take place, increasing passenger and crew comfort and safety.

Figure 4 shows another example where the multi-model ensemble approach is better, as the MOGREPS-G ensemble does not forecast the shear turbulence event but the ECMWF-EPS ensemble does. This case study is interesting however because the turbulence event is over the Rocky Mountains, so this could be MWT. It is probably a combination of shear and MWT and again shows the multi-model ensemble approach was a benefit, as if we only had the MOGREPS-G ensemble we would not forecast this event. This case also reinforces the need to have a MWT diagnostic, as the severe turbulence observations spread further than the forecast indicated. Figure 4 also shows some of the problems with turbulence forecasting, as we see what could be a false alarm event over Canada. Both the ECMWF and MOGREPS-G ensemble predict turbulence; however, there is no turbulence observed. This shows the benefit of using a probabilistic forecast because different end users can select the probability threshold of when they would take action. For example, if the probability forecast for turbulence is 20 % then there is a 1 in 5 chance of turbulence being observed. However if the

end user sets their threshold for action at 10 % then this event would be classed as a false alarm because it was forecasted (as it was above the 10 % threshold) and the event did not occur. If the threshold they used was 30 % however, this would be a correct rejection as turbulence would not be forecast as the threshold was not exceeded and turbulence did not occur. This helps to illustrate the trade-off between hits and false alarms and the ability the probability framework gives to its users to fine tune their response to optimise the forecast. So for this example, a higher threshold would result in a correct rejection, but might also miss the MoG turbulence event over the Rocky Mountains.

After plotting all of the 424 MoG turbulence events, we identified 98 cases that are likely to be CIT, which TI1 cannot forecast. To address this issue the CIT events from the rest of our study are removed in order to give the fairest possible test for the multi-model ensemble forecasts. We decided to keep all other MoG turbulence events in the study because we have no strong evidence that they are not shear related. An example is MWT, although we can identify these events occur over mountains we are not able to prove it is not shear turbulence. Using the satellite convection product we are more confident of the CIT events and therefore we have more confidence in removing them. If we are unsure in any way that it is not a CIT event, it is kept in the study. What this does highlight though is this study must be extended in the future to include a convective diagnostic and a MWT diagnostic because combined they count for a third of the events in this study.

After removing the events that we have categorised as CIT only it was possible to analyse the performance of the multi-model ensemble and the single-model ensembles and compare it to the previous studies such as Gill and Buchanan [2014]. Figure 5 is a ROC plot showing the skill score for both the single-model ensembles and the combined multi-model ensemble. Typically the area under the curve is a good measure of discriminatory skill. However, in this study the MOGREPS-G ensemble has a shorter line than the both the ECMWF and multi-

model ensemble. This is because a 12 member ensemble can't forecast the same lower probabilities as a larger ensemble. The 12 member ensemble can only predict probabilities as small as 1/12, whereas the ECMWF 51 member ensemble can forecast probabilities as low as 1/51. Therefore a simple Area Under the Curve (AUC) number could be biased towards the ECMWF forecast and multi-model ensemble forecasts as the longer line could (and does in this example) give them a larger AUC. Therefore it is better to focus on how steep the line is, and therefore on low false alarm rates that are more useful for the aviation industry, although the best method of measuring statistical significance is to use the AUC.

Low false alarm rates are more important for this study because airline companies may have a limit on acceptable hit rates and false alarm rates, and therefore the lower false alarm rates are the ones they would focus on. Figure 5 shows that the ECMWF, MOGREPS-G, and simple combined ensemble have almost the same skill. This is surprising because by combining the two ensembles, the forecast spread has increased and therefore we capture more turbulence events, but consequently more false alarms. Because of this trade off, we do not see a significant increase in skill. The AUC for the two single model ensembles are: ECMWF – 0.7712 and MOGREPS – 0.6881. The multi-model ensemble has an AUC of 0.7842 with the 95 % confidence interval lower bound being 0.7538 and the 95 % confidence interval upper bound being 0.8102. This therefore shows at the 95 % confidence interval, that the multi-model ensemble is only significantly better than the MOGREPS single model ensemble but not the ECMWF single model ensemble and this is because the MOGREPS-G line is shorter. To understand the benefit of using a multi-model ensemble Table 4 shows the number of MoG turbulence events where both models agree, and if they disagree, which model forecasted the turbulence event. Out of the 326 MoG turbulence events, 243 of them the models are in agreement, so they either both forecast the event, or both do not forecast the event. That leaves 83 out of 326 MoG turbulence when the models do not agree. What we

find however is this number is not split evenly and ECMWF forecasts 73 times when MOGREPS does not and there are only 10 occasions where MOGREPS forecasts turbulence and ECMWF does not. This suggests that there are only 10 occasions where it is a benefit to have a multi-model ensemble over a single ECMWF ensemble. This could be part of the reason why we do not see the large improvement if forecast skill when combining ensembles. Most of the events are already forecast by one of the models, and therefore we only have a limited benefit to adding the second ensemble.

The relative economic value of the forecast is shown in Figure 6 and it shows the multi-model ensemble has greater value than the single model ensembles but only for low cost/loss ratios. This is important because depending on the relative importance of minimising misses and maximising hits for an airline company, defines the cost/loss ratio we focus on. We also know the cost of action is likely to be a great deal less than the cost of loss due to injuries or aircraft damage. Therefore the lower cost/loss ratios are likely to be more important for the airline companies, and therefore this study focuses on those here. So the multi-model ensemble is as skilful as the single model ensembles, but would be more useful for decision-making in an operational environment. This figure also shows the maximum value for all the thresholds of the combined multi-model ensemble. As said before, the probability fields for many thresholds have been calculated, so this curve takes the highest value threshold for each cost/loss ratio. Since this bold line is above the others, there is more value in some of the other thresholds, and an optimised multi-model ensemble would provide more value and is worth pursuing in future studies. It is important to point out again that the ECMWF EPS value is theoretical and operationally would be lower since the availability of the data forces the use of a longer lead time. Also when comparing the single model ensembles to the previous study by Gill and Buchanan [2014] we see the improvement in the relative economic value, showing a significant model improvement over the last few years. This

improvement could be because the Met Office introduced the ENDGame (Even Newer Dynamics for General atmospheric modelling of the environment) dynamical core [Walters et al., 2014]. This has provided a better forecast and in particular has resulted in a 'reduction of the slow bias in tropospheric windspeeds'. It is important to note that a direct comparison is not possible because each study looks at two different years, but this study shows a large improvement in MOGREPS-G value compared to the Gill and Buchanan [2014].

Also plotted is a reliability plot shown in Figure 7. This figure shows that the MOGREPS-G, ECMWF-EPS, and combined multi-model ensembles under-forecast the lower probabilities, but over-forecast the higher probabilities. This is shown by each ensemble being above the line of a perfect forecast for the lower probabilities, but below the line for higher forecast probabilities. It is important to note that these plots have been calibrated because the forecast percentage from the ensembles are much higher than the observed frequency. This linear calibration is the forecast probability multiplied by a constant, which for this study is 1/17, which brings it more in-line with the observed frequency. Although a direct comparison can't be made, the forecast percentages and observed frequency in this example has increased and the reliability has improved over the last few years compared to Gill and Buchanan [2014]. What this again suggests is the turbulence forecast models have improved over the last few years and the multi-model ensemble is at least as reliable as the individual ensembles.

So far it has been shown that combining two ensembles improves the forecast, but it is also important to understand how the individual ensembles compare to each other. To do this we must first reduce the ensemble size of the ECMWF forecast to make it a fair comparison. This is because a larger ensemble should give a larger forecast spread, and therefore improve the forecast result, and the ECMWF EPS has 51 members compared to the MOGREPS-G 12 members. To do this we choose the first 12 members of the ECMWF ensemble. Each of the

perturbed members are constructed to be equally likely, and each consecutive member has a 'pair-wise anti-symmetric perturbation' [Owens and Hewson, 2018]. Therefore choosing consecutive members is a bias-free method for creating a sub sample. This is also how Buizza and Palmer [1998] studied the impact of ensemble size on ensemble skill. They took pairs of perturbed members, so that each ensemble has pairs of members with the same positive and negative perturbation. Figure 8 is the ROC plot with the same ensemble size and can see that both models have almost the same forecast skill. When looking at Figure 9 we see that the ECMWF-EPS ensemble is more valuable which is interesting to note. This would be useful when trying to combine the two ensembles using a weighted scheme to get the best forecast skill. As the ECMWF-EPS forecast is more valuable a larger weight would be applied when creating the multi-model ensemble. But again in an operational system the ECMWF-EPS skill would be reduced as the longer lead time needed due to the time delay of the forecast would reduce the skill. So before an optimised weighted multi-model ensemble can be created, the ECMWF ensembles performance with the time delay would have to be analysed. This would then need to be extended to include all turbulence predictors to find the best multi-model ensemble forecast.

6. Conclusions and further work

This study has investigated the role ensemble forecasts have in aviation turbulence. By combining two ensembles to create a simple multi-model ensemble, we aimed to show improved forecast value and skill which could then be implemented operationally. To verify the forecasts, aircraft observations from a fleet of Boeing 747 and 777 aircraft are used and created a contingency table of results. From this the results are analysed to show the multi-model ensemble system is as skilful as a single model ensemble, which follows on from the work of Gill and Buchanan [2014] and Buchanan [2016].

The results found suggest the forecast skill for the simple equally weighted multi-model ensemble is at least as skilful as the single model ensembles. This lack of significant improvement in the forecast skill was not expected, but this could be because when increasing the forecast spread, we capture more turbulence events, and also more false alarms. We would have to optimise this trade off to maximise the forecast skill, but in this study we are unable to show a significant improvement. However the value of the forecast is improved for the multi-model ensemble particularly at low cost/loss ratios, which are more important for operational use. Therefore to see an improvement in value at low cost/loss ratios shows it is worth pursuing this multi-model approach as it would be more valuable in an operational setting. Our results also showed that the multi-model ensembles are as reliable as the single model ensembles, and therefore overall the multi-model ensembles are an improvement to single model ensembles. Through combining two ensembles we gain consistency, gives more operational resilience and create one authoritative forecast whilst maintaining skill and increasing value, which would be particularly important in operational use in the future by the WAFCs.

Throughout the analysis, it is also found that the Ellrod and Knapp [1992] TI1 predictor is good at forecasting shear turbulence particularly. However not all the shear-induced turbulence events are forecast, and therefore one or more shear turbulence diagnostic would be beneficial similar to Kim et al. [2015]. It would also be a good next step to include the Ellrod3 turbulence diagnostic from Sharman and Pearson [2017], as they showed its improved performance over other turbulence diagnostics and could be an easy step to improving the forecast skill. Also the MOGREPS-G ensemble is designed to be time lagged to create a 24 member ensemble, and this should also be investigated in further work [Met Office, 2017]. An alternative method for creating a probabilistic forecast would be a multi-diagnostic approach, rather than the traditional ensemble members approach. Kim et al.

[2018] showed that a multi-diagnostic approach using two numerical models (NOAA's Global Forecast System and Met Office's Unified Model) for creating a probability forecast had a far greater statistical performance than the current WAFC forecast and any single CAT diagnostic. This is another example of how probabilistic forecasting can improve forecast skill, but uses a different method to create it. It is also vital to add convective and mountain wave predictors in any further studies. This would then take into account Convectively Induced Turbulence (CIT) and Mountain Wave Turbulence (MWT) that are also leading causes of aviation turbulence and account for many injuries to passengers and crew. Also CIT and MWT predictors could benefit more from two ensembles than the Ellrod and Knapp [1992] TI1 forecast. This could be because how the models forecast the predictors could be very different, therefore the forecast spread would be much higher between models, making a multi-model ensemble superior to a single-model ensemble.

References

Atlas D, Metcalf J, Richter J, Gossard E. 1970. The birth of 'CAT' and microscale turbulence. *J. Atmos. Sci.* **27**(6): 903–913.

Bedka, K.M., Brunner, J., Dworak, R., Feltz, W., Otkin, J. and Greenwald, T., 2010: Objective satellite-based detection of overshooting tops using infrared window channel brightness temperature gradients. *J. Appl. Meteorol. Climatol*. **49**, 181-202.

Bishop CH, Etherton BJ, Majumdar SJ. 2001. Adaptive sampling with the ensemble transform kalman filter. part i: Theoretical aspects. *Mon. Weather Rev.* **129**(3): 420–436.

Bowler NE, Arribas A, Mylne KR, Robertson KB, Beare SE. 2008. The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorolog. Soc.* **134**(632): 703–722.

Buchanan P. 2016. Aviation Turbulence Ensemble Techniques. In: *Aviation Turbulence*, Springer, pp. 285–296.

Buizza R, Palmer TN. 1998 Impact of ensemble size on ensemble prediction. *Monthly Weather Review*. **126**(9):2503-18.

Chambers E. 1955. Clear air turbulence and civil jet operations. *Aeronaut J.* **59**(537): 613–628.

Ehrendorfer M, Murphy AH. 1988. Comparative evaluation of weather forecasting systems: Sufficiency, quality, and accuracy. *Monthly Weather Review*, **116**(9): 1757-1770.

Ellrod GP, Knapp DI. 1992. An objective clear-air turbulence forecasting technique: Verification and operational use. *Weather Forecast.* **7**(1): 150–165.

Ellrod GP, Knox JA. 2010. Improvements to an operational clear-air turbulence diagnostic index by addition of a divergence trend term. *Weather and Forecasting*. **25**(2):789-98.

Endlich RM. 1964. The mesoscale structure of some regions of clear-air turbulence. *J. App. Meteorol.* **3**(3): 261–276.

Francis PN, Batstone C. 2013. *Developing a satellite product to identify severe convective storms hazardous to aviation.* Satellite Applications Technical Memo 11

Gill PG. 2014. Objective verification of World Area Forecast Centre clear air turbulence forecasts. *Meteorol. Appl.* **21**(1): 3–11.

Gill PG. 2016. Aviation Turbulence Forecast Verification. In: *Aviation Turbulence*, Springer, pp. 261–283.

Gill PG, Buchanan P. 2014. An ensemble based turbulence forecasting system. *Meteorol. Appl.* **21**(1): 12–19.

ICAO. 2012. Guidance on the harmonized WAFS grids for cumulonimbus cloud, icing and turbulence forecasts. (Available at https://www.icao.int/safety/meteorology/WAFSOPSG/GuidanceMaterial/Forms/AllItems.aspx.)

Jolliffe IT, Stephenson DB. 2012. *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Kane T, Brown B, Bruintjes R. 1998. Characteristics of pilot reports of icing. In: *Preprints: 14th Conference on Probability and Statistics, 11âĂß16 January 1998, Phoenix, AZ.* American Meteorological Society: Boston, MA

Kim JH, Chun HY. 2011. Statistics and possible sources of aviation turbulence over South Korea. *J. App. Meteorol. and Clim.* **50**(2): 311–324.

Kim, J. H., Chan, W. N., Sridhar, B., & Sharman, R. D. 2015. Combined winds and turbulence prediction system for automated air-traffic management applications *J. App. Meteorol. and Clim,* **54**(4): 766-784.

Kim JH, Sharman R, Strahan M, Scheck JW, Bartholomew C, Cheung JC, Buchanan P, Gait N. 2018. Improvements in Non-Convective Aviation Turbulence Prediction for the

World Area Forecast System (WAFS). *Bulletin of the American Meteorological Society. doi:10.1175/BAMS-D-17-0117.1, in press*

Knox, J. A., McCann, D. W. and Williams, P. D. 2008. Application of the Lighthill–Ford theory of spontaneous imbalance to clear-air turbulence forecasting. *Journal of the Atmospheric Sciences* **65**(10): 3292–3304.

Koch SE, Dorian PB. 1988. A mesoscale gravity wave event observed during CCOPE. Part III: Wave environment and probable source mechanisms. *Mont. weather rev.* **116**(12): 2570–2592.

Krishnamurti TN, Kishtawal C, Zhang Z, LaRow T, Bachiochi D, Williford E, Gadgil S, Surendran S. 2000. Multimodel ensemble forecasts for weather and seasonal climate. *J. Clim.* **13**(23): 4196–4216.

Lilly DK. 1978. A severe downslope windstorm and aircraft turbulence event induced by a mountain wave. *J. Atmos. Sci.* **35**(1): 59–77.

Marlton, G. J., Giles Harrison, R.., Nicoll, K. A., Williams, P. D. 2015. A balloon-borne accelerometer technique for measuring atmospheric turbulence. *Review of Scientific Instruments* **86**(1): 016109.

Met Office. 2017. MOGREPS-G guide to data (Available at https://www.metoffice.gov.uk/services/data-provision/big-data-drive/wholesale/categories/mogrepsg-user-guide)

Met Office. 2018. WAFC London Performance Indicators (Available at https://www.metoffice.gov.uk/public/weather/aviation-wafc/#?tab=wafcPerformance)

McCann, D. W., Knox, J. A. and Williams, P. D. 2012. An improvement in clear-air turbulence forecasting based on spontaneous imbalance theory: the ULTURB algorithm. *Meteorological Applications* **19**(1): 71–78.

Molteni F, Buizza R, Palmer TN, Petroliagis T. 1996. The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**(529): 73–119.

Owens, R G, Hewson, T D. 2018. *ECMWF Forecast User Guide*. Reading: ECMWF. doi: 10.21957/m1cs7h

Richardson DS. 2000. Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **126**(563): 649–667.

Schmetz, J., Tjemkes, S. A., Gube, M. and van de Berg, L., 1997: Monitoring deep convection and convective overshooting with METEOSAT, *Adv. Space Res*. **19**(3): 433-441

Schwartz B. 1996. The quantitative use of pireps in developing aviation weather guidance products. *Weather Forecast.* **11**(3): 372–384.

Sharman R, Lane T. 2016. *Aviation Turbulence: Processes, Detection, Prediction*. Springer.

Sharman R, Tebaldi C, Wiener G, Wolff J. 2006. An integrated approach to mid-and upper-level turbulence forecasting. *Weather Forecast.* **21**(3): 268–287.

Sharman R, Cornman L, Meymaris G, Pearson J, Farrar T. 2014. Description and derived climatologies of automated in situ eddy-dissipation-rate reports of atmospheric turbulence. *Journal of Applied Meteorology and Climatology.* **53**(6): 1416–1432.

Sharman RD, Pearson JM. 2017. Prediction of energy dissipation rates for aviation turbulence. Part I: Forecasting nonconvective turbulence. *Journal of Applied Meteorology and Climatology*. **56**(2):317-37.

Storer, L. N., Williams, P. D. and Joshi, M. M. 2017. Global response of clear-air turbulence to climate change. *Geophysical Research Letters.* **44**.

Storer, L. N., Williams, P. D. and Gill, P. G. 2018. Aviation Turbulence: Dynamics, Forecasting, and Response to Climate Change. *Pure Appl. Geophys*. https://doi.org/10.1007/s00024-018-1822-0

Swinbank R, Kyouda M, Buchanan P, Froude L, Hamill TM, Hewson TD, Keller JH, Matsueda M, Methven J, Pappenberger F, *et al.* 2016. The TIGGE project and its achievements. *Bull. Am. Meteorol. Soc.* **97**(1): 49–67.

Tebaldi C, Nychka D, Brown B G, Sharman R. 2002. Flexible discriminant techniques for forecasting clear-air turbulence. *Environmetrics*, **13**, 859-878.

Tenenbaum J. 1991. Jet stream winds: Comparisons of analyses with independent aircraft data over southwest Asia. *Weather Forecast.* **6**(3): 320–336.

Titley H, Stretton R. 2016. Tropical Cyclone Ensemble Forecasting at the Met Office: Upgrades to the MOGREPS Model and TC Products, and an Evaluation of the Benefit of Multi-model Ensembles . In: *Preprints: 32nd Conference on Hurricanes and Tropical Meteorology 17 22 April 2016, San Juan, PR*. American Meteorological Society: Boston MA.

Truscott B. 2000. EUMETNET AMDAR AAA AMDAR Software Developments Technical Specification. *Doc. Ref. E_AMDAR/TSC/003. Met Office: Exeter, UK* .

Uccellini LW, Koch SE. 1987. The synoptic setting and possible energy sources for mesoscale wave disturbances. *Mont. Weather Rev.* **115**(3): 721–729.

Vitart F. 2006. Seasonal forecasting of tropical storm frequency using a multi-model ensemble. *Q. J. R. Meteorol. Soc.* **132**(615): 647–666.

Walters D, Wood N, Vosper S, Milton S. 2014. *ENDGame: A new dynamical core for seamless atmospheric prediction*. Met Office documentation. (Available at http://www. metoffice.gov.uk/media/pdf/s/h/ENDGameGOVSciv2.0.pdf.)

Williams P. D. 2016. Transatlantic flight times and climate change. *Environmental Research Letters* **11**(2): 024008.

Williams PD. 2017. Increased light, moderate, and severe clear-air turbulence in response to climate change. *Adv. Atmos. Sci.* **34**(5): 576–586.

Williams, P. D., Read, P. L. and Haine, T. W. N. 2003. Spontaneous generation and impact of inertia–gravity waves in a stratified, two-layer shear flow. *Geophysical Research Letters.* **30**(24): 2255.

Willaims, P. D., Haine, T. W. N. and Read, P. L. 2005. On the generation mechanisms of short-scale unbalanced modes in rotating two-layer flows with vertical shear. *Journal of Fluid Mechanics.* **528**(11): 1–22.

Williams, P. D., Haine, T. W. and Read, P. L. 2008. Inertia gravity waves emitted from balanced flow: Observations, properties, and consequences. *Journal of the Atmospheric Sciences*, **65**: 3543–3556

WMO. 2003. Aircraft meteorological data relay (AMDAR) reference manual.

Captions

Figure 1: Plot of the spatial coverage of flight data from the fleet of Boeing 747 and 777 aircraft in May 2016.

Figure 2: Plot of a moderate-or-greater turbulence event over the possible sources of turbulence: top left: orography, shear turbulence (bottom left: MOGREPS-G and bottom right: ECMWF EPS probability forecast), and top right: convection from satellite data (colour shading indicates deep convection). Both the MOGREPS-G and ECMWF-EPS ensembles forecast the shear turbulence event. The circles indicate turbulence observations with grey indicating no turbulence, orange indicating light turbulence and red indicating moderate or greater turbulence. The convective classification can be found in Francis and Batstone [2013].

Figure 3: Plot of a moderate-or-greater turbulence event over the possible sources of turbulence: top left: orography, shear turbulence (bottom left: MOGREPS-G and bottom right: ECMWF EPS probability forecast), and top right: convection from satellite data (colour

shading indicates deep convection). Only the MOGREPS-G ensemble forecast the shear turbulence event. The circles indicate turbulence observations with grey indicating no turbulence, orange indicating light turbulence and red indicating moderate or greater turbulence. The convective classification can be found in Francis and Batstone [2013].

Figure 4: Plot of a moderate-or-greater turbulence event over the possible sources of turbulence: top left: orography, shear turbulence (bottom left: MOGREPS-G and bottom right: ECMWF EPS probability forecast), and top right: convection from satellite data (colour shading indicates deep convection). Only the ECMWF-EPS ensemble forecasts the shear turbulence event. The circles indicate turbulence observations with grey indicating no turbulence, orange indicating light turbulence and red indicating moderate or greater turbulence. The convective classification can be found in Francis and Batstone [2013].

Figure 5: ROC plot of the global turbulence with the 98 convective turbulence cases removed showing the forecast skill of the MOGREPS-G (dot-dash) AUC=0.6881, ECMWF (dot) AUC=0.772 and combined multi-model ensemble (dash) AUC=0.7842. The data used has a forecast lead time between +24 hours and +33 hours between May 2016 and April 2017.

Figure 6: Value plot with a log scale x-axis of the global turbulence with the 98 convective turbulence cases removed showing the forecast skill of the MOGREPS-G (dot-dash), ECMWF (dot), combined multi-model ensemble (dash) and the maximum value using every threshold of the combined multi-model ensemble (solid). The data used has a forecast lead time between +24 hours and +33 hours between May 2016 and April 2017.

Figure 7: Reliability diagram of the MOGREPS-G (dot-dash), ECMWF (dot) and combined multi-model ensemble (dash). The data used has a forecast lead time between +24 hours and +33 hours between May 2016 and April 2017.

Figure 8: ROC plot of the global turbulence showing the forecast skill of the MOGREPS-G (dot-dash) and ECMWF 12 member ensemble (dot). The data used has a forecast lead time between +24 hours and +33 hours between May 2016 and April 2017.

Figure 9: Value plot with a log scale x-axis of the global turbulence showing the forecast value of the MOGREPS-G (dot-dash) and ECMWF 12 member ensemble (dot). The data

used has a forecast lead time between +24 hours and +33 hours between May 2016 and April 2017.

Table 1: Turbulence severity for values of Derived Equivalent Vertical Gust (DEVG) [Truscott, 2000] and Eddy Dissipation Rate (EDR) [Sharman, 2014]. For severe turbulence to be observed the DEVG value must be greater than or equal to 9 m s-1 and therefore $9 \leq$ DEVG.

Table 2: A 2 × 2 contingency table showing the possible results of a turbulence forecast or event. The four possible outcomes include a Hit, Miss, False alarm and Correct Rejection.

Table 3: A 2 × 2 contingency table assigning a cost to the possible results of a turbulence forecast or event. The four possible outcomes include a Hit (with a subsequent cost), Miss (with a subsequent cost), False alarm (with a subsequent cost) and Correct Rejection (with no cost as no action was taken).

Table 4: Categorising moderate or greater turbulence events between cases where both ECMWF and MOGREPS models are in agreement (both do/do not forecast turbulence), and where the models are not in agreement (one model does forecast turbulence and the other does not). When the models are not in agreement, the results are put into a sub category of which ensemble did forecast the turbulence event.

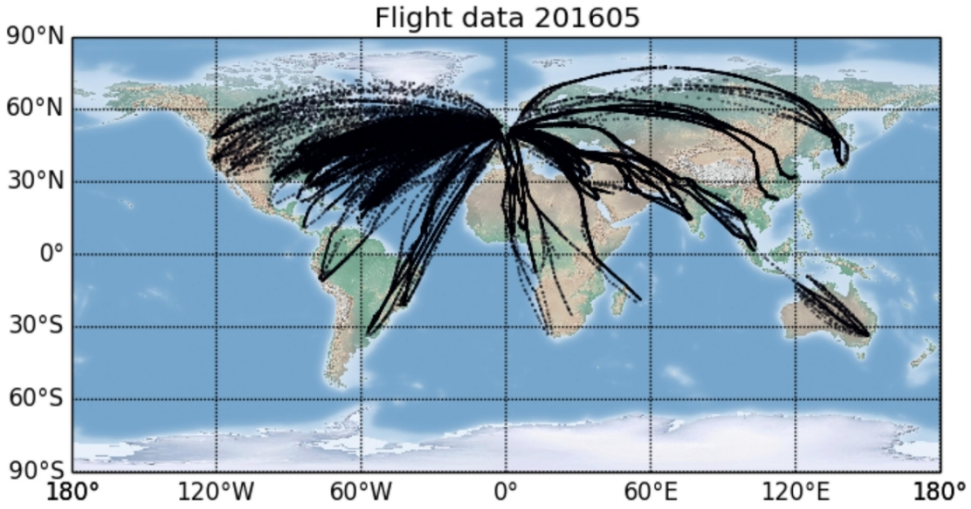

Figure 1: Plot of the spatial coverage of flight data from the fleet of Boeing 747 and 777 aircraft in May 2016.

103x59mm (300 x 300 DPI)

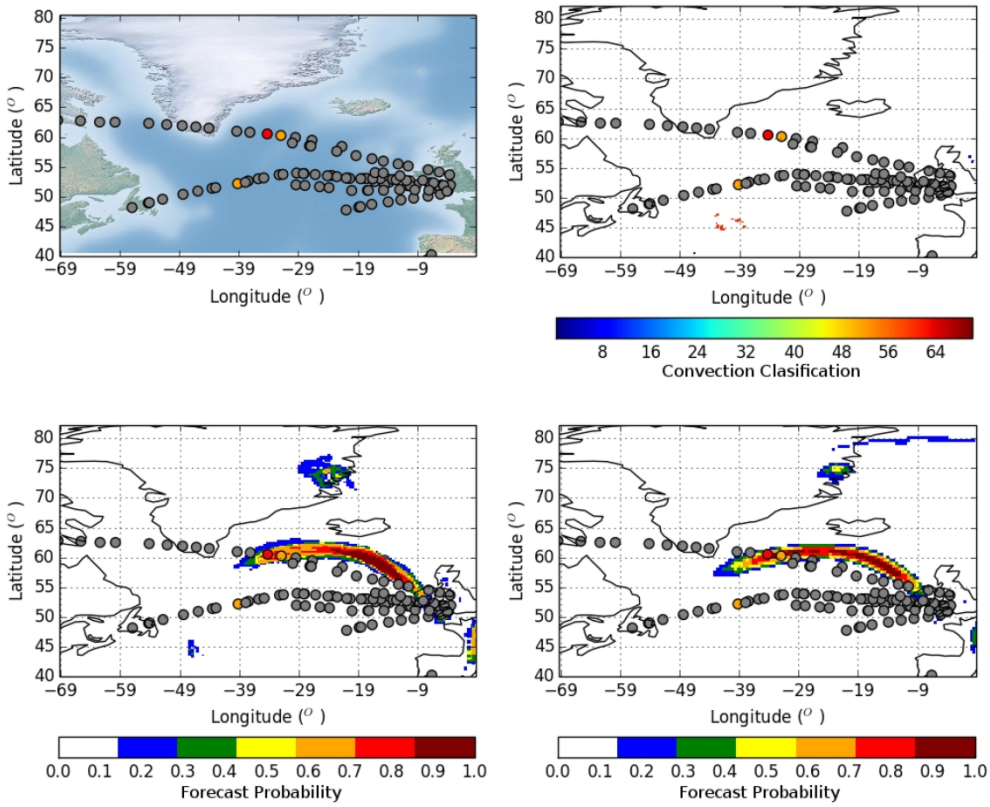

Figure 2: Plot of a moderate-or-greater turbulence event over the possible sources of turbulence: top left: orography, shear turbulence (bottom left: MOGREPS-G and bottom right: ECMWF EPS probability forecast), and top right: convection from satellite data (colour shading indicates deep convection). Both the MOGREPS-G and ECMWF-EPS ensembles forecast the shear turbulence event. The circles indicate turbulence observations with grey indicating no turbulence, orange indicating light turbulence and red indicating moderate or greater turbulence. The convective classification can be found in Francis and Batstone [2013].

213x177mm (300 x 300 DPI)

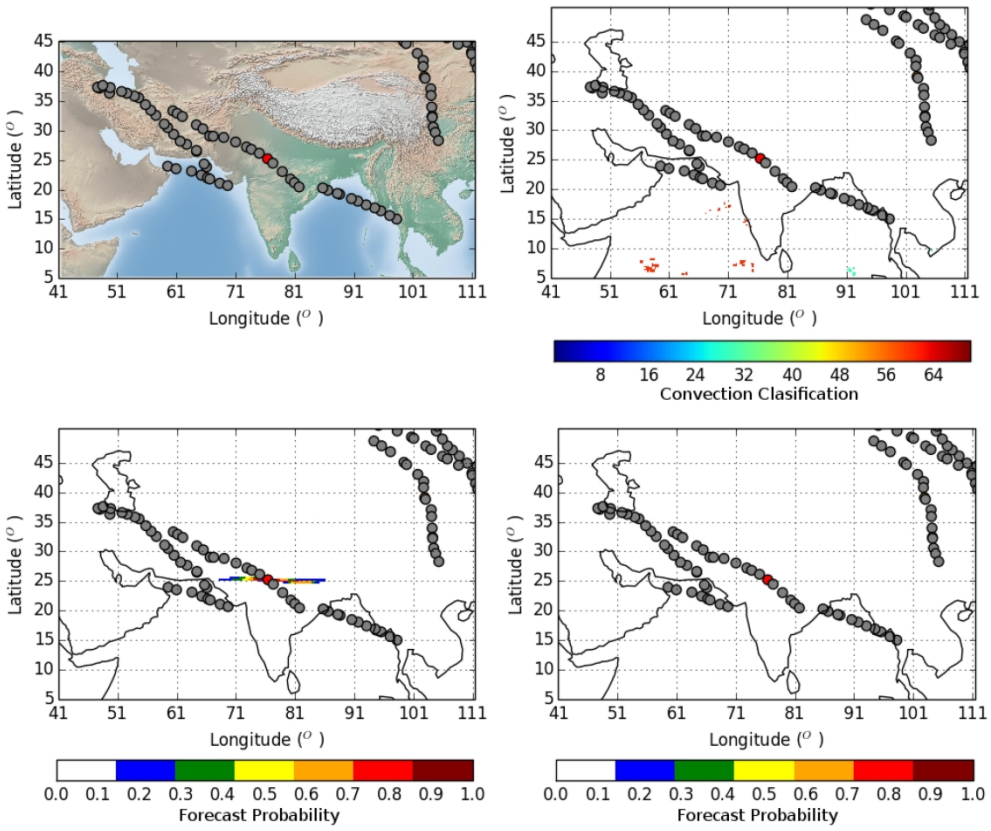

Figure 3: Plot of a moderate-or-greater turbulence event over the possible sources of turbulence: top left: orography, shear turbulence (bottom left: MOGREPS-G and bottom right: ECMWF EPS probability forecast), and top right: convection from satellite data (colour shading indicates deep convection). Only the MOGREPS-G ensemble forecast the shear turbulence event. The circles indicate turbulence observations with grey indicating no turbulence, orange indicating light turbulence and red indicating moderate or greater turbulence. The convective classification can be found in Francis and Batstone [2013].

213x177mm (300 x 300 DPI)

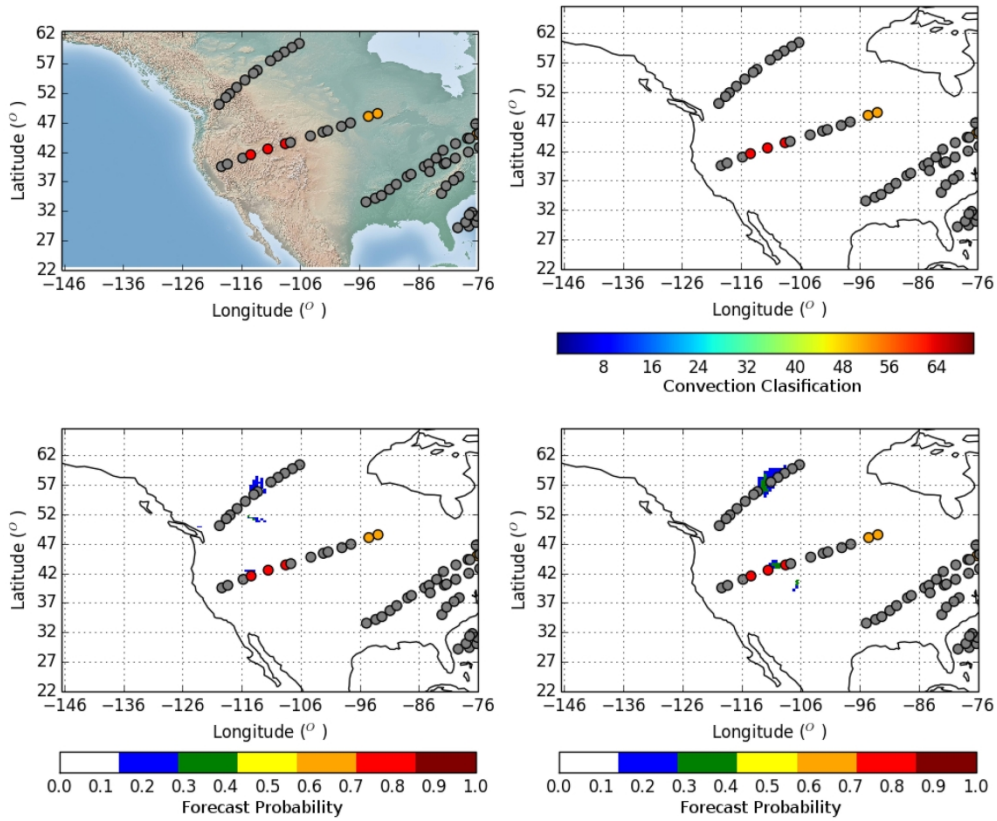

Figure 4: Plot of a moderate-or-greater turbulence event over the possible sources of turbulence: top left: orography, shear turbulence (bottom left: MOGREPS-G and bottom right: ECMWF EPS probability forecast), and top right: convection from satellite data (colour shading indicates deep convection). Only the ECMWF-EPS ensemble forecasts the shear turbulence event. The circles indicate turbulence observations with grey indicating no turbulence, orange indicating light turbulence and red indicating moderate or greater turbulence. The convective classification can be found in Francis and Batstone [2013].

213x177mm (300 x 300 DPI)

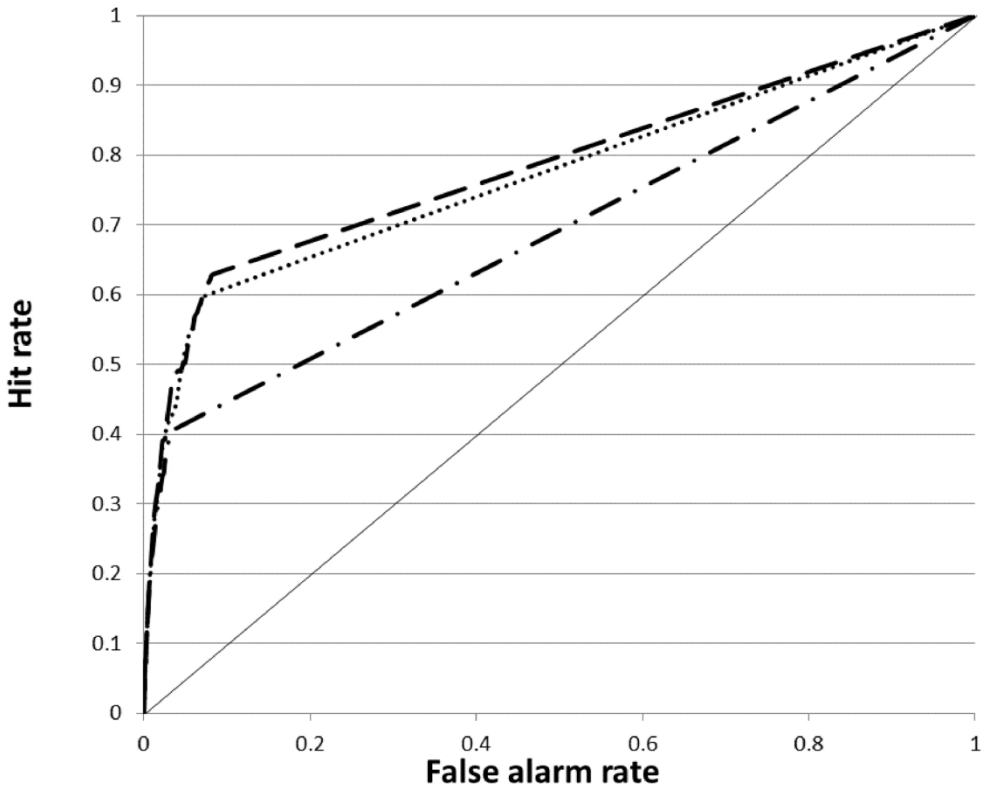

Figure 5: ROC plot of the global turbulence with the 98 convective turbulence cases removed showing the forecast skill of the MOGREPS-G (dot-dash) AUC=0.6881, ECMWF (dot) AUC=0.772 and combined multi-model ensemble (dash) AUC=0.7842. The data used has a forecast lead time between +24 hours and +33 hours between May 2016 and April 2017.

237x189mm (300 x 300 DPI)

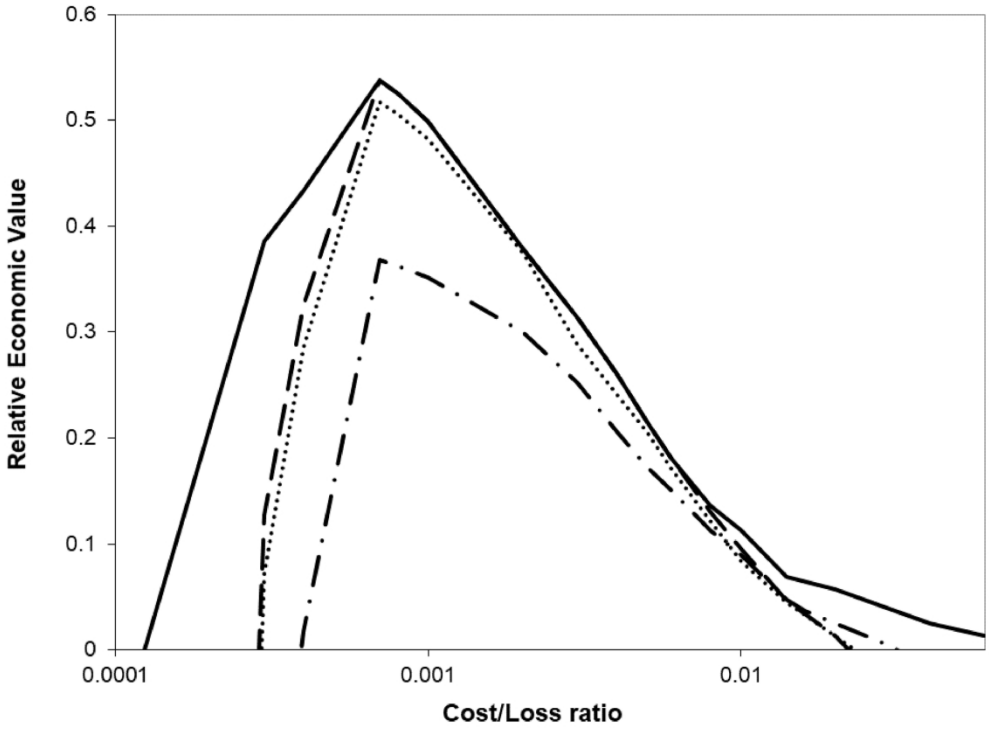

Figure 6: Value plot with a log scale x-axis of the global turbulence with the 98 convective turbulence cases removed showing the forecast skill of the MOGREPS-G (dot-dash), ECMWF (dot), combined multi-model ensemble (dash) and the maximum value using every threshold of the combined multi-model ensemble (solid). The data used has a forecast lead time between +24 hours and +33 hours between May 2016 and April 2017.

256x188mm (300 x 300 DPI)

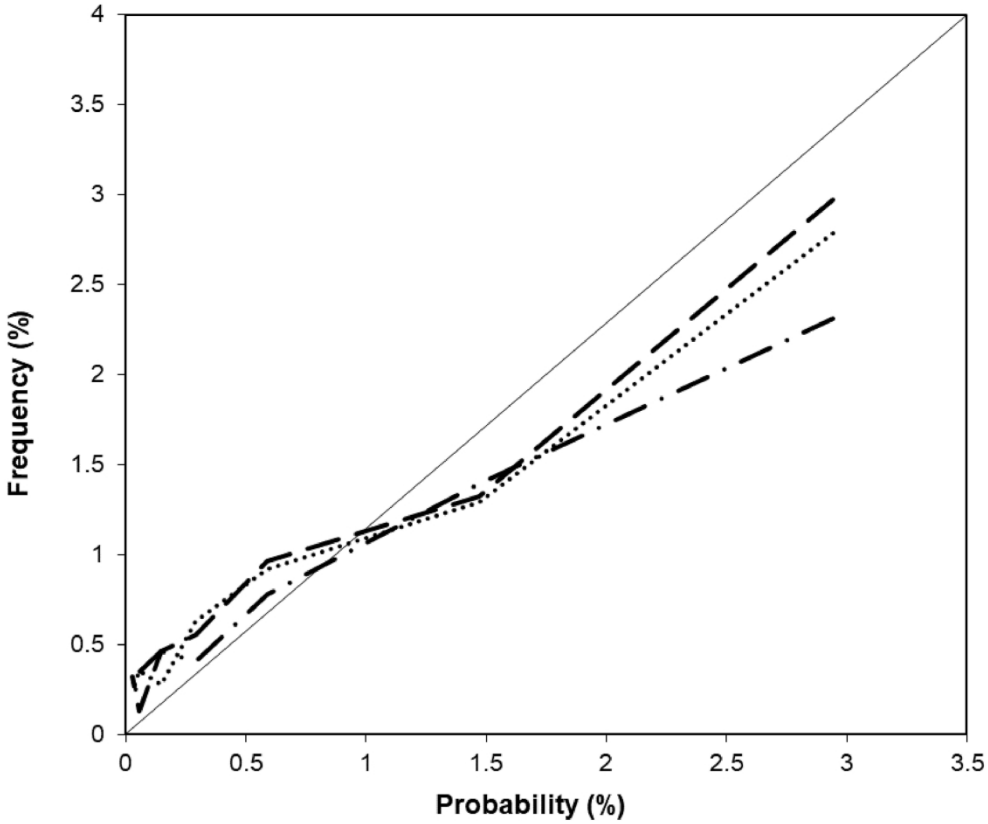

Figure 7: Reliability diagram of the MOGREPS-G (dot-dash), ECMWF (dot) and combined multi-model ensemble (dash). The data used has a forecast lead time between +24 hours and +33 hours between May 2016 and April 2017.

229x189mm (300 x 300 DPI)

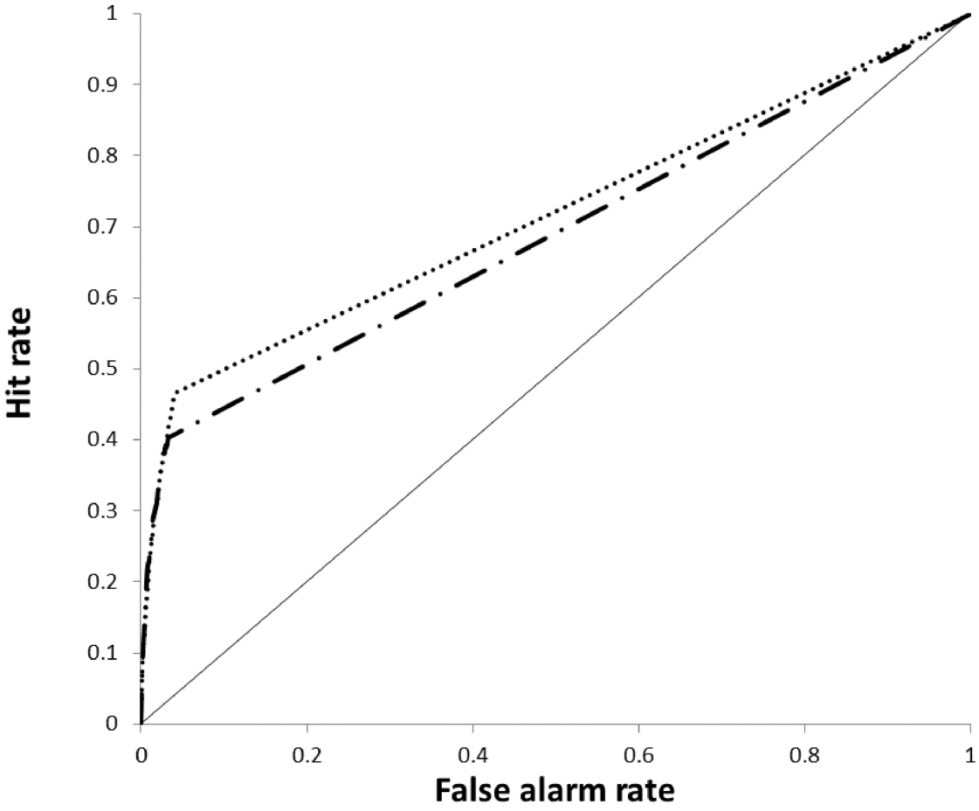

Figure 8: ROC plot of the global turbulence showing the forecast skill of the MOGREPS-G (dot-dash) and ECMWF 12 member ensemble (dot). The data used has a forecast lead time between +24 hours and +33 hours between May 2016 and April 2017.

231x190mm (300 x 300 DPI)

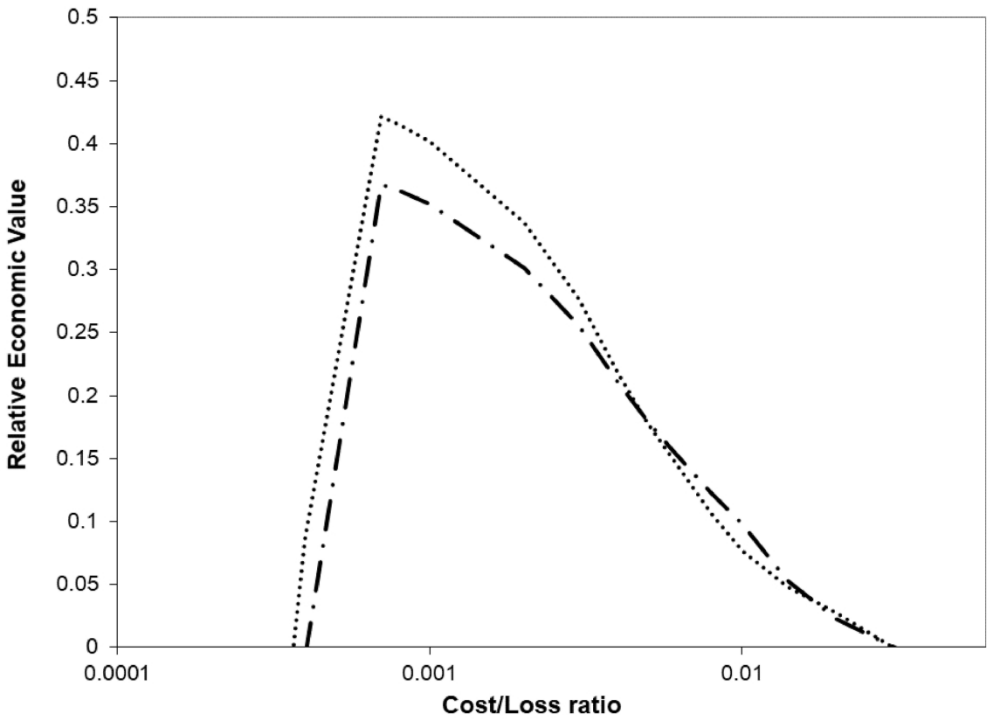

Figure 9: Value plot with a log scale x-axis of the global turbulence showing the forecast value of the MOGREPS-G (dot-dash) and ECMWF 12 member ensemble (dot). The data used has a forecast lead time between +24 hours and +33 hours between May 2016 and April 2017.

257x185mm (300 x 300 DPI)

Turbulence severity	**DEVG (m s^{-1})**	**EDR ($m^{2/3}$ s^{-1})**
None	DEVG ≤ 2	EDR ≤ 0.07
Light	2 ≤ DEVG ≤ 4.5	0.07 ≤ EDR ≤ 0.27
Moderate	4.5 ≤ DEVG ≤ 9	0.27 ≤ EDR ≤ 0.61
Severe	9 ≤ DEVG	0.61 ≤ EDR

	Turbulence observed	Turbulence not observed
Turbulence forecast	A (Hit)	B (False Alarm)
Turbulence not forecast	C (Miss)	D (Correct Rejection)

	Turbulence observed	Turbulence not observed
Turbulence forecast Action taken	Hit Cost	False alarm Cost
Turbulence not forecast No action taken	Miss Loss	Correct rejection

Models in agreement	Models not in agreement	
	ECMWF forecasts turbulence	MOGREPS forecasts turbulence
243	73	10