DTN Aviation Weather

Providing Weather intelligence for **safe** and **efficient** Flight Operations

The DTN Enhanced Flight Hazard Overview and Verification using In-situ EDR observations from Operational Commercial Aircraft

DTN

By Daniel W. Lennartson

1. Introduction

DTN delivers a numerical Enhanced Flight Hazard (EFH) turbulence product (Lennartson and McCann, 2014) that can be used for both flight planning (Strategic) as well as flight-following (Tactical) use. It is very important that an aviation turbulence forecast show reasonable accuracy spatially, temporally, and quantitatively to be a credible source for operational flight decision support. Shown in this document will be a volume evaluation of the EFH product to test its credibility in operational use by comparing it with turbulence observations and with verification of another well recognized numerical turbulence forecast, the Graphical Turbulence Guidance (GTG); Version 2.5, Sharman et al., 2006). The EFH and GTG approach turbulence forecasting very differently. The EFH is a deterministic forecast, and GTG is a weighted ensemble of diagnostic forecasts.

Verification was conducted by AvMet Applications, Inc. (AvMet). An analysis of EFH and GTG forecasts was conducted using data for a three-month period from August 2014 to October 2014.

A further analysis showing the core differences between DTN's physical approach versus GTG's statistical approach is also highlighted and features the reasoning behind the desired usage of the Lighthill-Ford method to capture CAT events. An interesting comparison between the methods that are also part of the statistical GTG ensemble of diagnostics shows their performances individually.

In the following sections, the DTN EFH forecast system will be briefly described; the verification methodology explained; the verification results and analysis shown; and finally, the conclusions from both the AvMet verification and the additional analysis of DTN EFH versus GTG forecasts.

2. Description of the EFH turbulence forecast

The EFH numerical turbulence forecast is a deterministic forecast derived from a numerical weather prediction model. The forecast is model agnostic and focuses on four primary sources of turbulence: mountain wave, boundary layer, upper-level clear air, and convective turbulence. Output from all modes is integrated into one eddy dissipation rate (EDR; Cornman et al., 1995) value, the rate at which turbulent energy dissipates into the atmosphere. Figure 1 shows conceptually how the forecast turbulence sources can change throughout a flight path and where they can potentially be enhanced where two or more sources are present at a given point at a given altitude.

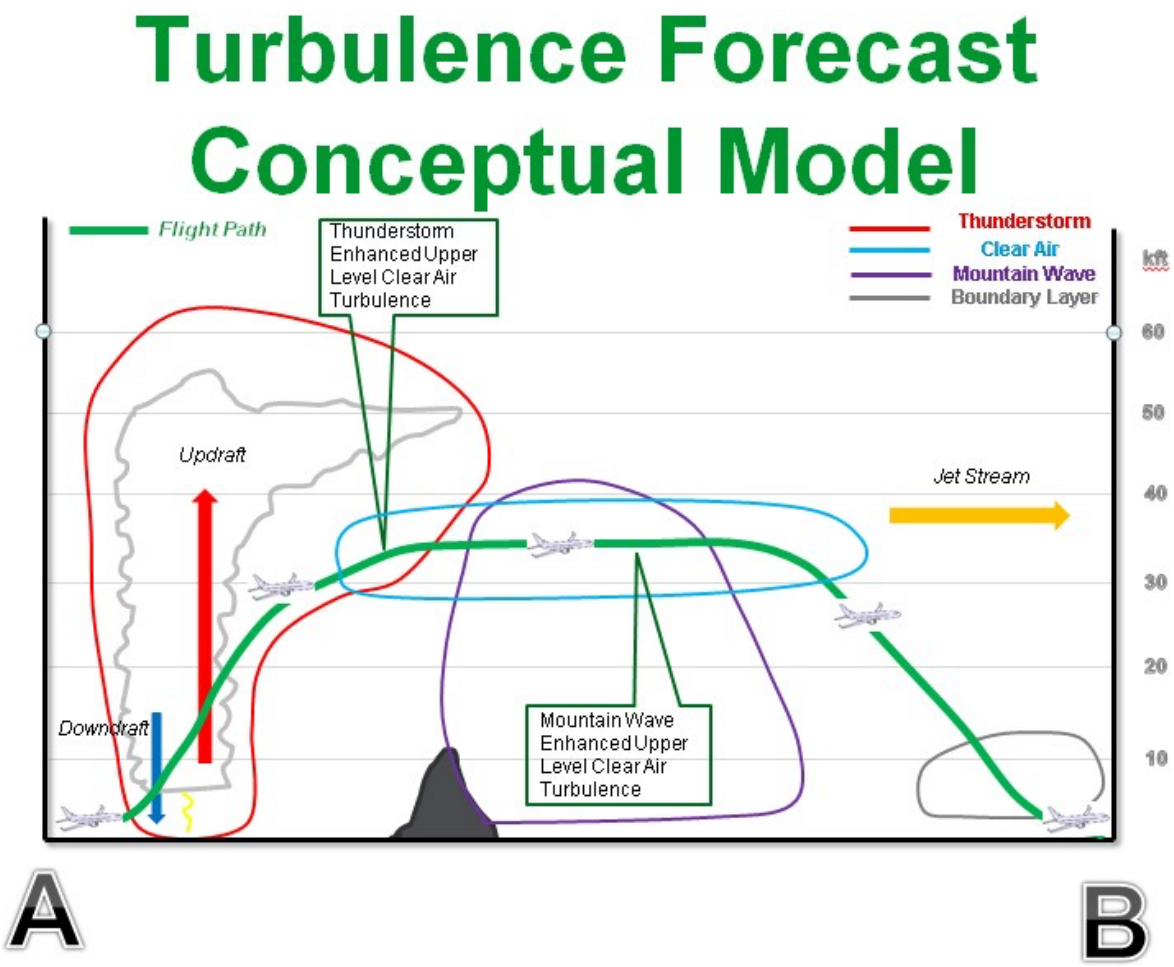


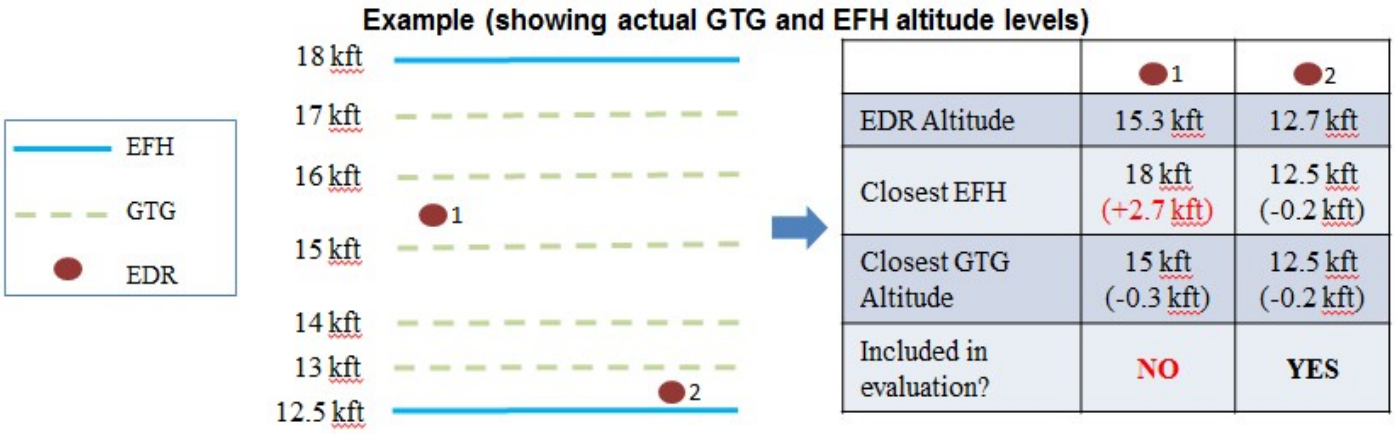
Figure 1: *Conceptual model featuring the four modes of the turbulence forecast: mountain wave, boundary layer, upper-level clear air, and convective. Also illustrated is the integration of the modes where constructive interference can enhance turbulence.*

2.1 Mountain Wave

The mountain wave component of the turbulence forecast calculates the turbulence potential of waves breaking over high terrain. This calculation takes into consideration the attributes of the mountain(s) such as asymmetry and concavity as well as the wind direction at the mountain top level. Also included are the effects of a hydraulic jump and the reflection/resonance of terrain induced mountain waves.

2.2 Boundary Layer

Boundary-layer turbulence results from the interaction of the lower atmosphere with the earth's surface. The turbulence values are calculated from the surface to the variable from-point-to-point top of the boundary layer defined as where boundary-layer EDR becomes zero (McCann, 2001).



	1	2
EDR Altitude	15.3 kft	12.7 kft
Closest EFH	18 kft (+2.7 kft)	12.5 kft (-0.2 kft)
Closest GTG Altitude	15 kft (-0.3 kft)	12.5 kft (-0.2 kft)
Included in evaluation?	NO	YES

Figure 2: *Observation 1 is an example of a rejected sample and observation 2 is an example of an accepted sample. The solid light blue lines represent EFH levels and dashed light green lines represent GTG levels.*

2.3 Upper-Level Clear-Air Turbulence

Upper-level Clear-Air Turbulence (CAT) is computed by applying Lighthill-Ford spontaneous imbalance theory to identify gravity waves that locally alter the environment's wind shear and stability. The altered state may be enough to lower the Richardson number to less than 0.25 thus initiating Kelvin-Helmholtz instability (Knox et al., 2008).

2.4 Convective

The convective turbulence component computes the turbulence related to vertical motions within convective clouds. This component is proportional to the updraft/downdraft strength (Byers and Braham, 1949). Furthermore, there are two additional thunderstorm features that are taken into account in the turbulence computation, gravity waves emitting outward from storm updrafts and the mountain wavelike turbulence associated with overshooting thunderstorm tops (Lennartson and McCann, 2014).

2.5 Outputs

The outputs offered for customer use are grib2, geotiffs, shapefiles, and geojsons. The typical data flow starts with the ingest of input models. The model data is then processed into the format required by the EFH algorithm processing. The EFH processing delivers the output and pushes the various output formats to a storage accessible to the various distribution methods we currently support (API, FTP, etc…). We currently have output available for 2 regions: Global at 0.25 deg and North America at 13km.

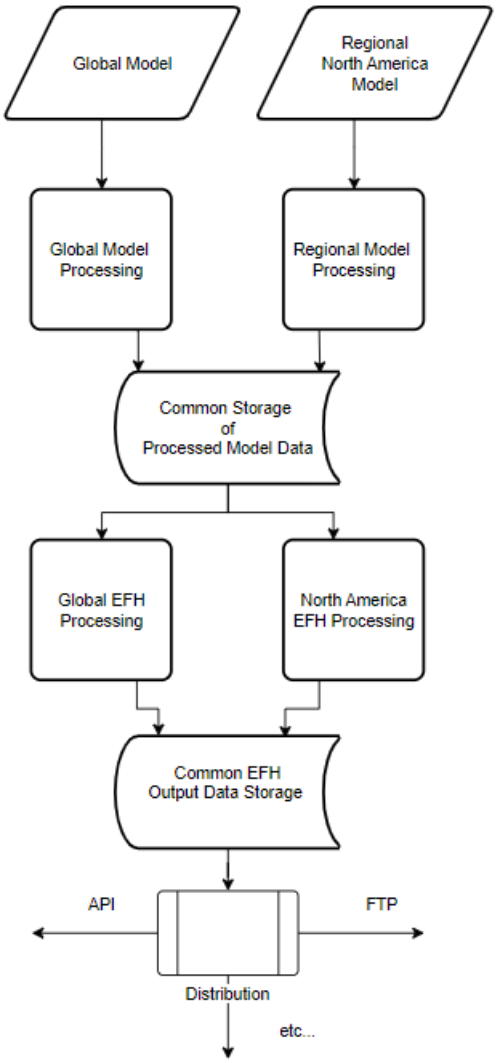


Figure 3: *EFH Data Flow Diagram showing the data flow from model ingest to EFH output distribution.*

3 Verification Methodology

Forecasts were validated against 121,576 EDR turbulence observations from commercial airlines over a span of 68 days from August 2014 to October 2014. The forecasts were validated over the common EFH and GTG forecast lead times of 1, 2, 3, 6, 9 and 12 hours. Sampling from the forecasts was defined horizontally as the average EDR forecast value within a 50 mile radius of each observation. Vertically, samples were defined as a mutual forecast level from EFH and GTG within 1 kft of the observation (Figure 2).

All null values for both products were considered as EDR values of 0 forecast (i.e., no turbulence), and both forecast datasets were treated as direct representations of observed in-situ EDR values.

4 Verification Approach and Definitions

A combination of three methods to analyze the forecasts are used: a contingency table showing the correlation between forecast and observed EDR values, graphs showing daily statistics, and summary tables showing all relevant statistics from both forecast models.

The contingency table shows counts of forecast EDR versus observed EDR in discrete bins (Figure 4). Ideally, forecasts and observations are perfectly correlated, and all data lie along the diagonal of the table. Off-diagonal elements indicate conditional biases in the forecasts.

1 Hour Lead Time
[Aug-Oct 2014]

Report EDR	Forecast EDR										
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	56209	9633	4051	1578	442	77	19	1	0	0	0
0.1	3672	2151	1187	410	111	26	8	0	0	0	0
0.2	322	237	153	64	18	10	0	0	0	0	0
0.3	32	25	22	10	2	0	0	0	0	0	0
0.4	4	6	3	2	0	0	0	0	0	0	0
0.5	0	1	0	0	0	0	0	0	0	0	0
0.6	0	0	0	0	0	0	0	0	0	0	0
0.7	0	0	0	0	0	0	0	0	0	0	0
0.8	0	0	0	0	0	0	0	0	0	0	0
0.9	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0

Figure 4: *Ranges of forecast turbulence versus observed EDR values where dark green boxes are perfect hits, red boxes are misses, yellow boxes are under-forecasts, and oranges are over-forecasts.*

Cells in the table are classified into four categories: "perfect" hits, under-forecasts, over-forecasts and misses. The respective definitions for each box represented in Figure 4, and the color coding applied to each, are given below:

Condition for Perfect Hit (dark green):
Case 1:
(Observed EDR > 0.1) AND
(Forecast EDR – Observed EDR ≤ 0.1)

Case 2:
(Observed EDR ≤ 0.1) AND
(Forecast EDR ≤ 0.1)
Condition for Over-forecast Hit (orange):
(Observed EDR > 0.1) AND
(Forecast EDR – Observed EDR ≥ 0.2)

Condition for Under-forecast Hit (yellow):
(Observed EDR > 0.3) AND
Forecast EDR > 0.1) AND
(Observed EDR – Forecast EDR ≤ 0.1)

Condition for False Alarm (red):
(Observed EDR > 0.1) AND
(Forecast EDR ≤ 0.1)

From the data in the table, summary statistics can be derived to measure performance such as: perfect hit rate, over-forecast rate, underforecast rate, overall hit rate, false alarm rate, and false alarm ratio and are defined below:

Perfect Hit Rate = #Perfect Hits/#EDR Observations

Overforecasted Hit Rate = #Overforecast Hits/ #EDR Observations

Underforecasted Hit Rate = #Underforecasted Hits/ #EDR Observations

Hit Rate = #Hits/#EDR Observations

#Hits = #Perfect Hits + #Overforecasted Hits + #Underforecasted Hits

False Alarm Rate = #False Alarms/#Observed EDR <= 0.1

False Alarm Ratio = #False Alarms/#Forecasted EDR > 0.1

Unforecasted Turbulence Rate = #Unforecasted Turbulence Occurrences/ #EDR Observations

#Unforecasted Forecast Occurrences = #Forecast EDR <= 0.1 AND #Observed EDR > 0.2

Summary of GTG vs DTN Data
[Aug-Oct 2014]

		Lead Hours					
		1	2	3	6	9	12
False Alarm Rate	GTG	33.4%	33.3%	32.1%	30.5%	29%	27.3%
	Schneider	21.9%	20.3%	19.4%	17.4%	16.2%	15.5%
False Alarm Ratio	GTG	80.1%	80.6%	80.6%	80.7%	81.4%	81.5%
	Schneider	78%	77.9%	77.8%	78.1%	78.2%	77.8%
Unforecasted Turbulence Rate	GTG	0.4%	0.4%	0.4%	0.5%	0.5%	0.6%
	Schneider	0.4%	0.5%	0.5%	0.6%	0.6%	0.6%
Perfect Hit Rate	GTG	68.4%	68.5%	69.8%	71.4%	72.9%	74.5%
	Schneider	79.2%	80.6%	81.4%	83.2%	84.3%	84.9%
Overforecasted Hit Rate	GTG	1.2%	1.2%	1%	0.8%	0.6%	0.5%
	Schneider	0.7%	0.7%	0.7%	0.6%	0.5%	0.6%
Underforecasted Hit Rate	GTG	0%	0%	0%	0.1%	0%	0%
	Schneider	0%	0.1%	0%	0%	0%	0%
Hit Rate	GTG	69.7%	69.7%	70.8%	72.2%	73.5%	75%
	Schneider	79.9%	81.4%	82.1%	83.8%	84.9%	85.5%
EDR Observations	GTG	80486	81105	82512	86114	84664	82356
	Schneider	80486	81105	82512	86114	84664	82356

Figure 5: *Summary Statistics Table showing all relevant statistics to evaluate forecast quality. These statistics include all days from the evaluation period for EFH (DTN) and GTG.*

5 Summary of AvMet Verification Results

Data were analyzed in aggregate and in several subsets including by forecast lead time, EDR value, and proximity to convection.

Figure 5 in section 4, shows in green that. EFH has the advantage over GTG consistently for false alarm rate and ratio and also in perfect hits and overall hits when evaluating the entire period. The reason the hits are as elevated as they are in large part because ~85% of the observations used in the evaluation are zero. So, in this evaluation period, forecasting EDR below 0.1 can make a big difference in the overall statistics.

a)

DTN
2 Hour Lead Time
[Aug-Oct 2014]

Report EDR	Forecast EDR						
	0	0.1	0.2	0.3	0.4	0.5	0.6
0	57947	9166	3727	1365	370	79	11
0.1	3864	2082	1000	392	127	20	3
0.2	357	235	140	57	30	5	0
0.3	34	30	24	10	8	0	0
0.4	5	9	3	3	0	0	0
0.5	0	1	0	0	0	0	0

b)

GTG
2 Hour Lead Time
[Aug-Oct 2014]

Report EDR	Forecast EDR						
	0	0.1	0.2	0.3	0.4	0.5	0.6
0	48435	15379	6741	1917	190	4	0
0.1	2275	2361	1902	869	78	3	0
0.2	290	240	182	95	17	0	0
0.3	40	29	26	11	0	0	0
0.4	10	3	5	1	1	0	0
0.5	0	0	0	1	0	0	0

Figure 6*: 2 Hour Lead Time Range of Forecast and Observation EDR for a) EFH and b) GTG.*

In Figures 6a and 6b, EFH has ~9,000 more Perfect Hit nulls than GTG. To put it in perspective, there were only ~8,500 observations of ≥ 0.1 EDR so forecasting EDR < 0.1 and Observed EDR ≥ 0.1, this number is far less than when there was Forecast EDR ≥ 0.1 and Observed EDR = 0 (i.e., a miss). That shows us both models tended to significantly over-forecast.

Also noteworthy is the benefit EFH has by having a convective algorithm included in its turbulence forecast suite. When categorized as non-convective versus convective (i.e., convection within 50 miles), the statistics heavily favor EFH when convection is present and the forecast EDR ≥ 0.2 (Figure 7). Proximity to convection was determined by identifying National Convective Weather Diagnostic (NCWD) VIP Level 3-or-greater cells within 50 miles of an EDR observation.

a)

Summary of GTG vs DTN
[Aug-Oct 2014; Convection within 50 miles]

		Lead Hours					
		1	2	3	6	9	12
False Alarm Rate	GTG	41.1%	40.9%	39.4%	36.6%	33.1%	31.1%
	DTN	46.8%	42.3%	38.6%	33.2%	27.3%	26.8%
False Alarm Ratio	GTG	67.6%	68.1%	68.3%	69.1%	71.3%	70.2%
	DTN	72.4%	71.2%	71.5%	71.1%	71.8%	70.9%
Unforecasted Turbulence Rate	GTG	1.5%	1.6%	1.8%	2%	2.3%	2.2%
	DTN	1.2%	1.4%	1.5%	1.9%	2.1%	1.9%
Perfect Hit Rate	GTG	65.1%	65.3%	66.8%	68.7%	71.1%	72.8%
	DTN	60.5%	64%	66.8%	71.1%	75.6%	75.8%
Overforecasted Hit Rate	GTG	1.5%	1.4%	0.8%	0.9%	0.6%	0.7%
	DTN	2%	1.9%	1.7%	1.3%	0.9%	1.2%
Underforecasted Hit Rate	GTG	0.2%	0.3%	0.3%	0.2%	0.1%	0.1%
	DTN	0.2%	0.3%	0.2%	0.1%	0.1%	0.1%
Hit Rate	GTG	66.7%	67%	67.9%	69.8%	71.9%	73.6%
	DTN	62.6%	66.2%	68.7%	72.6%	76.6%	77.2%
EDR Observations	GTG	8002	8093	8263	8404	8243	8293
	DTN	8002	8093	8263	8404	8243	8293

b)

Summary of GTG vs DTN
[Aug-Oct 2014; Convection within 50 miles; EDR Greater Than 0.2]

		Lead Hours					
		1	2	3	6	9	12
Unforecasted Turbulence Rate	GTG	42.1%	41.4%	44.0%	51.5%	63.8%	59.4%
	DTN	33.7%	35.4%	38.3%	48.8%	59.0%	52.8%
Perfect Hit Rate	GTG	51.9%	50.5%	49.1%	42.0%	32.4%	37.0%
	DTN	58.2%	53.0%	51.8%	45.4%	35.8%	41.3%
Overforecasted Hit Rate	GTG	1.1%	1.6%	0.0%	0.3%	0.0%	0.0%
	DTN	2.5%	3.8%	4.5%	2.5%	2.0%	2.3%
Underforecasted Hit Rate	GTG	4.9%	6.6%	6.9%	6.2%	3.8%	3.6%
	DTN	5.6%	7.8%	5.4%	3.4%	3.1%	3.6%
Hit Rate	GTG	57.9%	58.6%	56.0%	48.5%	36.2%	40.6%
	DTN	66.3%	64.6%	61.7%	51.2%	41.0%	47.2%
EDR Observations	GTG	285	319	334	324	293	303
	DTN	285	319	334	324	293	303

Figure 7: *Summary Statistics Table showing all relevant statistics where forecast EDR > 0.2. Rows in green show where EFH has the advantage a) All Days near Convection and b) All Days near Convection with Forecast EDR ≥0.2.*

As mentioned previously in this section, ~8,500 have Observed EDR ≥ 0.1. The overall scores for that subset show the Perfect Hits that GTG accumulates is directly a result of over-forecasting because there was a greater coverage of Forecast EDR ≥ 0.1 than EFH. The net effect makes the GTG Perfect Hit scores higher, but they were higher at the cost of a higher False Alarm Rate. When analyzing the observed EDR ≥ 0.3 EDR hits, EFH higher statistical numbers show its higher precision with elevated turbulence values (Figure 8a-c) at this threshold. Note sample sizes are relatively large for EDR ≥ .1 and .2, however decrease significantly for EDR ≥.3.

a)

Summary of GTG vs DTN
[Aug-Oct 2014; EDR Greater Than 0.1]

		Lead Hours					
		1	2	3	6	9	12
Unforecasted Turbulence Rate	GTG	3.6%	4.0%	4.1%	4.5%	5.1%	5.4%
	DTN	4.2%	4.7%	4.8%	5.7%	6.2%	6.1%
Perfect Hit Rate	GTG	84.2%	84.1%	85.7%	87.5%	89.1%	89.3%
	DTN	88.5%	88.0%	88.2%	88.0%	88.7%	88.0%
Overforecasted Hit Rate	GTG	11.8%	11.5%	9.7%	7.5%	5.4%	4.9%
	DTN	6.9%	6.8%	6.6%	6.0%	4.8%	5.5%
Underforecasted Hit Rate	GTG	0.3%	0.5%	0.5%	0.5%	0.5%	0.4%
	DTN	0.4%	0.5%	0.4%	0.3%	0.3%	0.3%
Hit Rate	GTG	96.4%	96.0%	95.9%	95.5%	94.9%	94.6%
	DTN	95.8%	95.3%	95.2%	94.3%	93.8%	93.9%
EDR Observations	GTG	8476	8439	8504	9022	8857	8620
	DTN	8476	8439	8504	9022	8857	8620

b)

Summary of GTG vs DTN
[Aug-Oct 2014; EDR Greater Than 0.2]

		Lead Hours					
		1	2	3	6	9	12
Unforecasted Turbulence Rate	GTG	33.8%	35.8%	36.9%	40.8%	46.4%	48.5%
	DTN	39.3%	41.6%	43.0%	51.8%	56.5%	55.6%
Perfect Hit Rate	GTG	62.0%	58.5%	57.6%	53.8%	49.1%	47.8%
	DTN	53.8%	50.2%	49.4%	43.7%	38.7%	39.1%
Overforecasted Hit Rate	GTG	1.3%	1.8%	1.2%	0.7%	0.3%	0.3%
	DTN	3.1%	3.7%	3.8%	2.1%	2.1%	2.3%
Underforecasted Hit Rate	GTG	2.9%	4.0%	4.3%	4.7%	4.1%	3.4%
	DTN	3.8%	4.5%	3.8%	2.4%	2.8%	2.9%
Hit Rate	GTG	66.2%	64.2%	63.1%	59.2%	53.6%	51.5%
	DTN	60.7%	58.4%	57.0%	48.2%	43.5%	44.4%
EDR Observations	GTG	911	951	944	995	967	953
	DTN	911	951	944	995	967	953

c)

Summary of GTG vs DTN
[Aug-Oct 2014; EDR Greater Than 0.3]

		Lead Hours					
		1	2	3	6	9	12
Unforecasted Turbulence Rate	GTG	40.2%	39.4%	38.1%	38.1%	51.2%	49.6%
	DTN	33.6%	30.7%	31.3%	47.5%	53.5%	45.7%
Perfect Hit Rate	GTG	35.5%	30.7%	31.3%	28.1%	17.8%	25.6%
	DTN	33.6%	35.4%	39.6%	34.5%	24.0%	31.8%
Overforecasted Hit Rate	GTG	0%	0%	0%	0%	0%	0%
	DTN	0.0%	0.0%	2.2%	0.7%	1.6%	0.8%
Underforecasted Hit Rate	GTG	24.3%	29.9%	30.6%	33.8%	31.0%	24.8%
	DTN	32.7%	33.9%	26.9%	17.3%	20.9%	21.7%
Hit Rate	GTG	59.8%	60.6%	61.9%	61.9%	48.8%	50.4%
	DTN	66.4%	69.3%	68.7%	52.5%	46.5%	54.3%
EDR Observations	GTG	107	127	134	139	129	129
	DTN	107	127	134	139	129	129

Figure 8: *Summary Statistics Table showing all relevant statistics where forecast EDR a)≥ 0.1, b)≥0.2, and c)≥0.3. Rows in green show where EFH has the advantage.*

When evaluating the results from the entire study period, the DTN EFH forecasts had higher Hit Rates, averaging 11% higher, overall forecast periods compared to the GTG forecasts. The False Alarm Rates for GTG were on average 14% higher than the EFH forecasts (Figure 9).

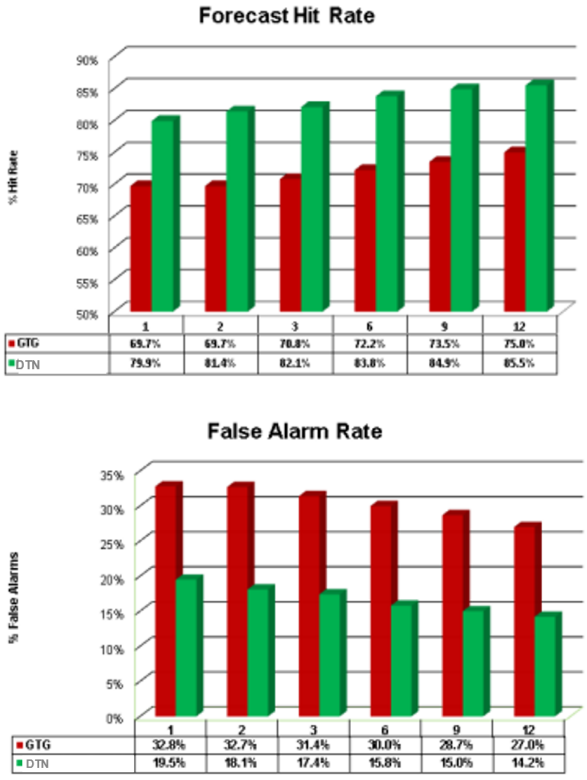


Figure 9: *Top graph shows the comparison of Hit Rate between EFH (green) and GTG (red). The bottom graph shows the False Alarm Rate between EFH (green) and GTG (red). Both graphs represent data from all lead times.*

Analyses of daily summary statistics over the evaluation period shows DTN EFH to be more consistent (i.e., less deviation from average) compared to GTG with the RMSE error consistently less than GTG through the evaluation period (Figure 10). This shows the EFH forecast technique to be a positive step forward in day-to-day consistency for the numerical prediction of turbulence.

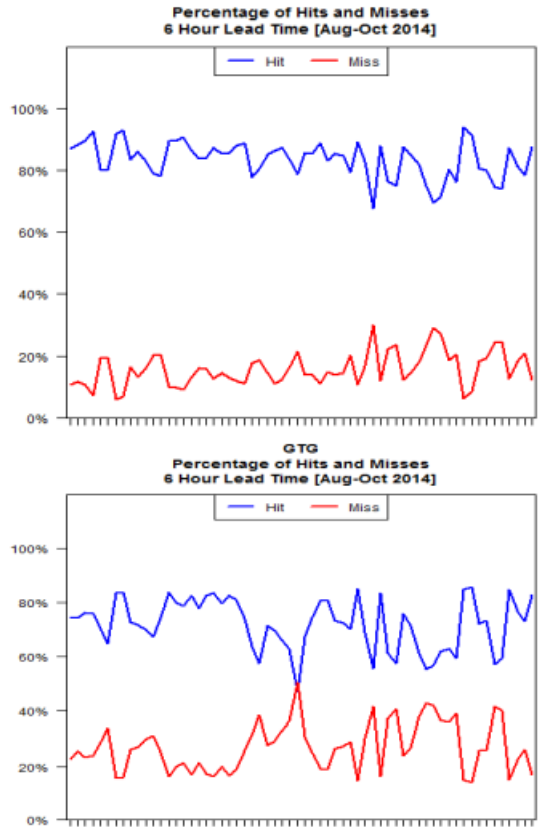


Figure 10: *Daily Summary of Results showing hits (blue) and misses (red) between EFH (top) and GTG (bottom) illustrating GTG's larger deviation from day to day than EFH.*

6 AvMet Verification Conclusions

DTN has a numerical deterministic turbulence forecast as a part of an Enhanced Flight Hazard product suite. To assess its potential operational merit, AvMet was tasked to conduct an independent third-party evaluation of the EFH turbulence forecast. The results from the three-month evaluation period of turbulence forecast information provided by EFH and GTG (v2.5) products as compared to EDR in-situ observations. Evaluations included identifying the accuracy of the products via contingency tables showing relationships between forecast and observed EDR values, graphs showing daily statistics, and summary tables presenting relevant statistics from both forecast models. Results from the evaluation showed:

- For all data, EFH has a consistently lower False Alarm Rate compared to GTG for all lead times evaluated (EFH = 18.5%, GTG = 30.9%)
- For all data, EFH scored higher for the 'perfect Hit Rate' and overall Hit Rate with ~11% improvement over GTG noted at all lead times for both statistics - EFH was also observed to have a more consistent validation (i.e., lower RMSE comparing to Hit Rate on average) for all forecast periods for all days evaluated compared to GTG
- EFH validated higher than GTG when there were larger turbulence observations (i.e., ≥.1, .2, and .3)
- EFH validated higher than GTG within vicinity (50 miles) of convection for all thresholds ≥ 0.1 EDR

The validation effort has shown the merit in EFH turbulence forecast data that offers opportunities for its end users to better optimize their routes thus conserving fuel, reducing emissions, and helps to reduce air traffic congestion. End users will also have the benefit of more actionable information with higher accuracy where dangerous conditions exist. Finally, the user of the EFH forecasts will have better confidence employing the forecast into operations with its smaller variations in quality from day to day. The difference in forecasting approaches between the deterministic EFH and ensemble-of-diagnostics GTG seems to have a lot to do with the deviations in results. For one, ensembles tend to spread variably from one weather pattern to the next since the ensemble members will agree and disagree variably. The result is more False Alarms from the broader coverage, dampening of highest values where divergent solutions can potentially cancel out each another, and more variability in day to day forecast confidence depending on daily member convergence or divergence. Also, the application of Lighthill-Ford theory as it is implemented in EFH turbulence model has good skill. As McCann et al. (2012) alluded to, advances in gravity wave initiation theories are an area that can be improved and could yield positive advances in quality on future numerical turbulence forecasting.

7 Physical versus Statistical Turbulence Forecast

When Knox et al. (2008) (hereafter KMW) computed the Lighthill-Ford diagnostic, it had to be converted into a non-dimensional gravity wave amplitude so it could be used as input into the turbulence model equations. How to make such a conversion has not been theoretically derived or observationally observed. However, Williams et al. (2008) noted that the gravity wave amplitudes in laboratory experiments vary linearly with Rossby number, and the Rossby number varies with the square root of the Lighthill-Ford radiation term. Therefore, KMW assumed the same relationship:

$$\hat{a} \propto \sqrt{R}$$

KMW empirically found a proportionality constant by examining the range of R in several CAT outbreaks, then, because the maximum â is 2.5, the constant was (max R)/2.5[2] .

For this experiment we assume that the other gravity wave diagnostics have a similar relationship. McCann (2001) gave value ranges for the divergence tendency. For frontogenesis, the Plougonven-Zhang diagnostic, stability advection, and acceleration divergence we found maximum values similarly as KMW found for Lighthill-Ford. As it turns out, computed EDRs were only somewhat sensitive to the constants we found if they were reasonably representative. Using the ULTURB software (McCann et al. 2012) as a template, we wrote similar programs for each of the other five gravity wave diagnostics, substituting the appropriate diagnostic for the Lighthill-Ford one.

We created a combined clear air and mountain wave GTG3 above FL200 (Sharman and Pearson 2017) replica forecast following their method. Because we only had archived verification data (see below), we had to be able to create forecasts from past model data. Our GTG3 version is not exact because the official GTG3 dynamically varies its weights for each input diagnostic. In our facsimile we weighted all input diagnostics equally as Sharman and Pearson suggest. While there are some differences, our version using the Rapid Refresh (RAP) forecast model looks very similar to official GTG3 forecasts as seen in Figure 12 and in other comparisons. We verified the physical gravity wave model using each of the six gravity wave diagnostics and our GTG3 forecast against automated in situ aircraft turbulence observations (Cornman et al. 1996). Onboard aircraft software analyzes aircraft movement to determine an aircraft-independent turbulence measurement. Observations can be transmitted as frequently as one per minute and include the average eddy dissipation rate (EDR) and the maximum EDR for the previous minute. Observations are archived on the National Centers for Environmental Prediction (NCEP) Meteorological Assimilation Data Ingest System (MADIS) website (www.madis.ncep.noaa.gov). We retrieved maximum EDR observations at or above FL200 plus or minus ten minutes of the top of every hour every day from 25 November 2017 to 4 March 2018. Our goal was to compare CAT gravity wave model forecasts and GTG3 with CAT observations. The chosen 100-day period climatologically provided the most CAT observations relative to convective induced turbulence observations. The maximum EDR observations were rounded to the nearest 0.1 m2/3 s-1 and put into bins of 0.0, 0.1, 0.2, 0.3, 0.4, and 0.5 or greater $m^{2/3}s^{-1}$. We averaged more than 21 000 observations per day distributed as in Figure 11. This distribution is very similar to other larger studies, i.e., Sharman et al. (2006).

EDR	Percentage
0.0	99.265
0.1	0.655
0.2	0.071
0.3	0.007
0.4	0.0009
≥0.5	0.0002

Figure 11: *Percentage distribution of EDR observations in each 0.1 m2/3 s-1 EDR bin*

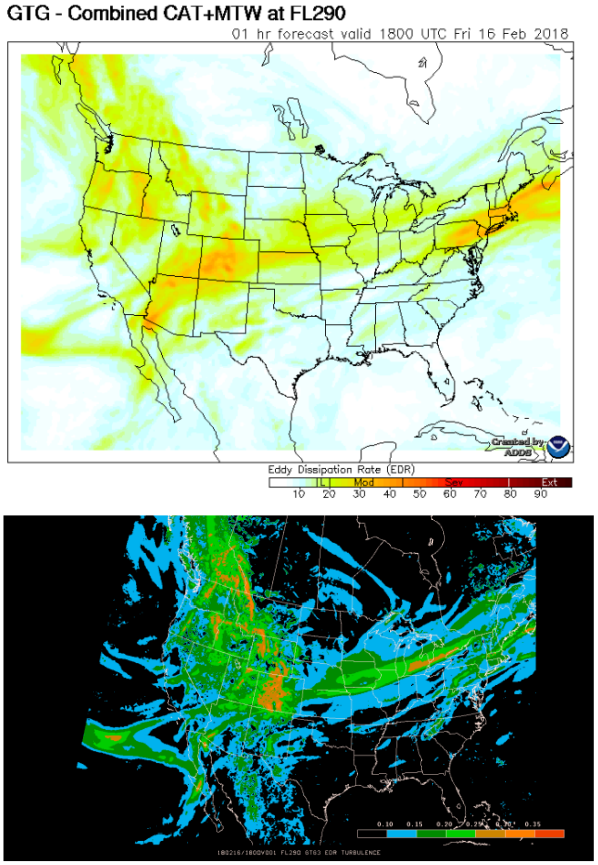


Figure 12: *(Top) One hour combined GTG3 forecast at fL290 valid 1800UTC 16 February 2018 taken from the Aviation Weather Center website (www.aviationweather.gov). (Bottom) Our GTG3 replica for the same forecast time and altitude.*

We matched the EDR observations with archived Rapid Refresh (RAP) numerical forecast model data retrieved from the National Centers for Environmental Information (www.ncdc.noaa.gov/nomads/dataproducts). The archived RAP model forecasts cover the contiguous United States, southern Canada, northern Mexico, and adjacent waters. The model runs every hour with archived forecasts available hourly on top of each hour out to 18 hours. We excluded any EDR observation outside the RAP model domain. For each first hour RAP model forecast we ran each of our gravity wave turbulence algorithms and our GTG3 version, all of which produced EDR forecasts. For each EDR observation ±10 minutes that forecast time we interpolated the forecast EDR to the exact location of the observation and rounded the forecast to the nearest 0.1 m $^{2/3}$ s^{-1}. We then binned them the same as the observations. We could pair each observation/forecast in a six-by-six contingency table as in Figure 13. Because we did not distinguish turbulence observations caused by CAT mechanisms, mountain waves, or convection, our analyses of the 6 X 6 contingency tables do not absolutely compute skill for CAT-only forecasts. But since our goal was to establish superiority of one method over another, our analyses are relative.

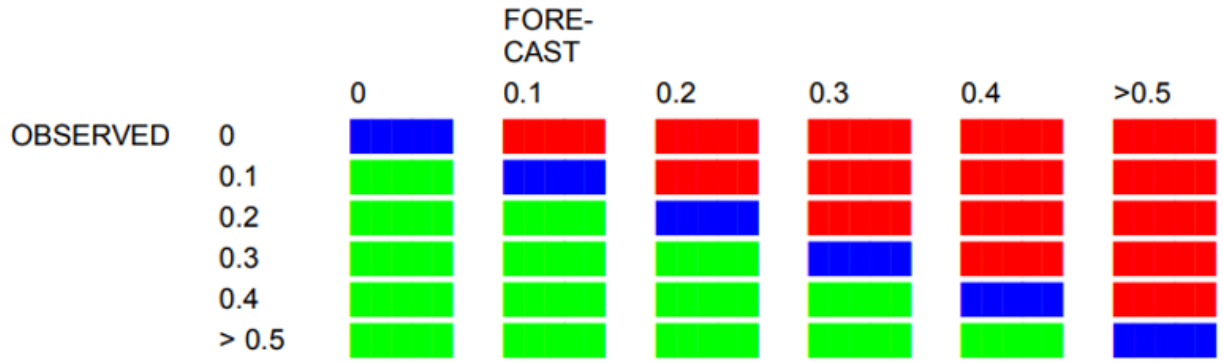


Figure 13: *Observed and forecast EDRs are paired and put into the appropriate place in the 6 X 6contingency table. Blue represents accurate forecasts, red represents overforecasts, and green represents underforecasts. The overall 6 X 6 contingency table may be easily reduced to smaller degrees such as 2 X 2 depending what statistics are desired.*

		FORECAST						
		0	0.1	0.2	0.3	0.4	>0.5	
OBSERVED	0	2078177	34525	23322	12586	4210	918	2153738
	0.1	11619	974	851	419	160	99	14122
	0.2	1180	100	102	83	44	33	1542
	0.3	115	13	9	6	8	3	154
	0.4	13	1	3	0	1	1	19
	>0.5	2	2	0	1	0	0	5
		2091106	35615	24287	13095	4423	1054	2169580

Figure 14: *The 6 X 6 verification contingency table for turbulence forecasts using frontogenesis.*

Diagnostic	PODsmooth	TSS
Lighthill-Ford	.849	.022
Divergence Tendency	.946	.016
Frontogenesis	.965	.034
Plougonven-Zhang	.928	.036
Stability Advection	.989	.072
Acceleration Divergence	.994	.078
GTG3	.474	.0076

Figure 15: *Probability of Detection of EDR< 0.05 $m^{2/3} s^{-1}$ (smooth) and overall True Skill Scores (TSS) for turbulence forecasts each of the six gravity wave diagnostics and for GTG3.*

First, we show how well each turbulence forecast fared overall. We calculated a 6 X 6 True Skill Score (Doswell et al. 1990)

$$TSS = \frac{tr(\mathbf{R})}{tr(\mathbf{R}^*)}$$

where R is a matrix of the contingency table values minus their expected value if the values were random and R* is a similar matrix of perfect forecast values. The tr function is the diagonal of the matrix (blue values in Fig. 13). For these contingency tables the TSS measures how well the forecasts are within 0.1

$m^{2/3}\ s^{-1}$ of the observations. These are stringent criteria, so because the EDR observations may not all be CAT-related, we are not looking for absolute values but are looking for relative skill of one method over another. The TSS measures how well the observations fall into the contingency table's diagonal (blue in Fig. 13). Figure 14 is the 6 X 6 contingency table for frontogenesis. Figure 15 shows the overall verification statistics for turbulence forecasts using each of the six gravity wave diagnostics and for GTG3. The highest TSS for acceleration divergence is about ten times the lowest TSS for GTG3. Even the low TSS for divergence tendency is more than twice GTG3's. There is a rough correlation between the Probability of Detection of EDR < 0.05 $m^{2/3}s^{-1}$ (hereafter "smooth") and the TSS. The fraction of the grid volume of forecast EDR to the total grid volume helps explain each forecast's TSS. This fraction is the number of grid points between and including FL200 and FL400 forecast above a certain threshold to the total number of grid points. Figure 16 shows these fractions. Given that more than 99% of the observations were smooth, it is not surprising that those methods that forecast small fractions of grid volume have the highest overall TSSs. That nearly half of GTG3's grid point volume has EDR > 0.05 $m^{2/3}\ s^{-1}$ explains why its overall TSS is so low. We generated bias statistics of our observations/forecasts. The bias is simply the number of forecasts divided by the number of observations above a certain threshold. Figure 17 shows these bias statistics. What immediately jumps out is the grossly overforecast of the turbulence by almost all methods, especially when forecasting EDR > 0.25 $m^{2/3}s^{-1}$. The highest biases are for Lighthill-Ford.

	EDR Forecast greater than				
Forecast	0.05	0.15	0.25	0.35	0.45
Lighthill-Ford	.130	.075	.030	.008	.0015
Divergence Tendency	.054	.039	.019	.003	.0003
Frontogenesis	.024	.017	.008	.002	.0005
Plougonven-Zhang	.068	.043	.013	.003	.0007
Stability Advection	.011	.005	.002	.0002	.00006
Acceleration Divergence	.007	.002	.0004	.00004	.000003
GTG3	.445	.053	.003	.00003	.000006

Figure 16: *The forecast fraction of grid points above various thresholds. This fraction is the number of grid points between and including FL200 and FL400 forecast above a certain threshold to the total number of grid points.*

	Bias for EDR Forecast greater than				
Forecast	0.05	0.15	0.25	0.35	0.45
Lighthill-Ford	10	79	318	716	518
Divergence Tendency	1.8	33	237	294	60
Frontogenesis	2.5	16	85	233	211
Plougonven-Zhang	8.0	19	106	171	187
Stability Advection	1.2	4.5	11	14	1.2
Acceleration Divergence	0.8	2.0	4.1	2.1	0.4
GTG3	71	90	64	12	0.2

Figure 17: *Bias statistics for the various gravity wave diagnostics and GTG3. The bias is the number of forecasts divided by the number of observations above a certain threshold.*

Forecast	POD(EDR>0.25)	POD(EDR<0.05)	TSS
Lighthill-Ford	.331	.849	.180
Divergence Tendency	.095	.946	.041
Frontogenesis	.112	.965	.077
Plougonven-Zhang	.179	.928	.107
Stability Advection	.051	.989	.040
Acceleration Divergence	.033	.994	.027
GTG3	.051	.474	-.475

Figure 18: *Probabilities of Detection of EDR > 0.25 $m^{2/3}$ s^{-1} and EDR < 0.05 $m^{2/3}$ s^{-1} and the resulting True Skill Score for each turbulence forecast method using automated aircraft EDR observations.*

While the overall TSSs mostly indicate how well the methods forecast smooth turbulence, users want to avoid the strong turbulence. To measure that skill, we reduced the 6 X 6 contingency tables to 2 X 2 tables setting an EDR > 0.25 $m^{2/3}s^{-1}$ threshold for PODyes and EDR < 0.05 $m^{2/3}s^{-1}$ for PODno. For a 2 X 2 table, the TSS is simply (PODyes + PODno – 1). Figure 18 shows the results. Here it is the methods with large volumes of high EDR that excel with the Lighthill-Ford method leading the way. Not only does GTG3 poorly forecast smooth turbulence, but it also doesn't forecast strong turbulence well thus yielding a very low negative skill. Additionally, we have been saving clear air turbulence case studies on which we have been testing ULTURB (McCann et al. 2012) which uses McCann's (2001) physical gravity wave model with the Lighthill-Ford gravity wave diagnostic. We have 41 cases within the RAP operational numerical forecast model domain above FL200 between 2012-2018 and within the Rapid Update Cycle model between 2010-2012. We gathered the cases from the archives of the Aviation Herald (www.aviationherald.com). The Aviation Herald reports on commercial aircraft mishaps worldwide, and the turbulence reports were primarily when the turbulence caused injuries. The Aviation Herald does not report on smaller general aviation aircraft, nor does it report every incident or accident. The Aviation Herald's editor verifies each report from two independent sources or from a government aviation safety agency. Of course, there were many more than 41 turbulence reports in the archive, so we confirmed that each report was clear air by eliminating reports associated with convection by examining satellite imagery. Because it can be difficult to distinguish mountain wave turbulence from clear air turbulence, we did not filter our data for the former. The Aviation Herald reports typically gave the event's time, location, and altitude. Even with that information, the incident was often reported when the aircrew took action rather than when the turbulence occurred. Whenever the report was inadequate in some way, we obtain the flight's track from FlightAware (www.flightaware.com).

We assumed the incident could have occurred as much as 10 minutes prior to the reported time. We discarded any report that we could not resolve to this accuracy. We verified each of the 41 cases similarly as we did for the automated EDR observations. If the EDR > 0.4 $m^{2/3}s^{-1}$, we considered it a "hit" for the tested diagnostic. EDR = 0.4 $m^{2/3}s^{-1}$ is about moderate to severe turbulence for most commercial aircraft (Sharman et al. 2014). We assume that the PODno statistic computed from our automated EDR observations is representative of each diagnostic's smooth turbulence forecasting skill in general.

Forecast	POD(EDR>0.40)	POD(EDR<0.05)	TSS
Lighthill-Ford	.829	.849	.678
Divergence Tendency	.366	.946	.312
Frontogenesis	.341	.965	.306
Plougonven-Zhang	.463	.928	.391
Stability Advection	.195	.989	.184
Acceleration Divergence	.122	.994	.116
GTG3	.171	.474	-.355

Figure 19: *Probabilities of Detection of EDR > 0.40 $m^{2/3} s^{-1}$ and EDR < 0.05 $m^{2/3} s^{-1}$ and the resulting True Skill Score for each turbulence forecast method using 41 cases of significant turbulence found in the Aviation Herald archives 2010-2018.*

Lighthill-Ford CAT forecast		
Statistic	Original	Reduced
TSS overall	.022	.052
POD(EDR<0.05 $m^{2/3} s^{-1}$)	.849	.952
TSS (EDR>0.25 $m^{2/3} s^{-1}$)	.180	.115
TSS (EDR>0.40 $m^{2/3} s^{-1}$)	.678	.464
Forecast Fraction (smooth)	.130	.039
Forecast Fraction (EDR>0.25 $m^{2/3} s^{-1}$)	.030	.0007

Figure 20: *Statistics comparing CAT forecasts from the original Lighthill-Ford diagnostic to one reduced by half.*

Figure 19 shows the results for each diagnostic. Four of the diagnostics captured more than 30% of the Aviation Herald reports indicating that the physical gravity wave model describes well the clear air turbulence production process. The Lighthill-Ford diagnostic is by far the best of the bunch, finding 83% of the reports. In fact, none of the other diagnostics captured any of the seven reports that Lighthill-Ford missed. Williams et al. (2013) suggested that some reported turbulence is wake turbulence caused by other nearby aircraft. Thus, some of the "misses" may be explained in this fashion. As an additional experiment, we halved the proportionality constant that converts the Lighthill-Ford diagnostic to non-dimensional gravity wave amplitude. Because of the assumed square root relationship between any diagnostic and gravity wave amplitude, this reduced the effect to a quarter. Figure 20 compares the original and the reduced Lighthill-Ford diagnostics. While the reduced Lighthill-Ford diagnostic forecast was better overall

with a smaller forecast fractions and smaller biases, it suffered when forecasting stronger turbulence. The results from this additional experiment and the overall statistics illustrate the ever-apparent tradeoff between smaller forecasts and forecasts of significant events. While the ideal forecast will be small and highlight the strong CAT, it is obvious that the research has not advanced far enough to date to do so. Furthermore, we would like to comment on the inadequacy of the GTG3 as a CAT forecast. From the forecast fractions in Figure 16, the GTG3 forecasts much more positive turbulence, nearly half the forecast volume, compared with any diagnostic used in the gravity wave initiating conceptual model presented in section 2. Moreover, GTG3's forecasts of significant turbulence, i.e., EDR > 0.35 $m^{2/3}s^{-1}$ are nearly non-existent. We conclude that GTG3 forecasts both ends of the turbulence spectrum poorly. We entered into this experiment with the idea of augmenting or changing our Lighthill-Ford diagnostic with another one if it were to improve our turbulence forecasts. As it turns out, we believe that we already have the best diagnostic and are distributing the most valuable clear air turbulence forecasts to our customers. These forecasts are available from DTN (https://www.dtn.com/industries/weather/aviation/).

8 Physical versus Statistical Analysis Conclusions

We presented a simple ingredients-based conceptual model of CAT. Gravity waves locally alter the environmental stability and wind shear when passing through. If the modification can lower the Richardson number to less than 0.25, then CAT will develop. The turbulence maximum potential intensity is computed from the modified stability and wind shear. Thus, we have three ingredients, environmental stability and environmental wind shear from which we compute an environmental Richardson number and gravity wave non-dimensional amplitudes. We compared six gravity wave indicators' turbulence forecast skill on automated aircraft EDR observations using the physical gravity wave conceptual CAT model. Because of the overwhelming number of smooth observations, overall, the acceleration divergence indicator, which forecast the smallest airspace volume, fared the best. In contrast, the Lighthill-Ford indicator, which forecast the largest airspace volume, did the best on stronger turbulence, EDR > 0.25 $m^{2/3}s^{-1}$. When we tested the gravity wave indicators on 41 significant turbulence events occurring between 2010-2018, the Lighthill-Ford indicator excelled over all the others with an 83% POD at EDR > 0.4 $m^{2/3}s^{-1}$. Because the Lighthill-Ford indicator was not too far behind the other gravity wave indicators in forecasting smooth, its True Skill Score, 0.678, was much higher than the others, all < 0.4. The statistically based GTG3 forecast paled in comparison to the physical gravity wave model forecasts at both ends of the turbulence spectrum. Its overall skill score was two to ten times lower than the gravity waves', depending on which method was the GTG3 compared, and its 17% POD for significant turbulence was near the bottom, resulting in a -0.355 TSS. Because of the high bias for forecasting positive CAT for the physical gravity wave conceptual model, especially the Lighthill-Ford version, one should not take its forecasts too literally. Users should interpret them as forecasting CAT potential, not actual CAT; an aircraft entering a high EDR volume is not likely to experience the maximum EDR forecast, but it could. This is analogous to tornado watches which cover vast areas compared to the tornadoes' actual areal extent. We cannot overemphasize how important the

physical gravity wave conceptual CAT model is to diagnose and forecast CAT. Without its framework, gravity wave indicators are not successful (e.g., Wilson 2012). Within its framework all of the gravity wave indicators that we tried work with a varying degree of success. We wouldn't be surprised if other more traditional CAT indices that may be related to gravity wave generation were applied in this simple model's context, they would have some success also. DTN recognizes that clear air, mountain wave, and convective turbulence may interact with each other. Therefore, they now integrate forecast algorithms of each (Lennartson and McCann 2014). Airlines and other users are discovering that they can proactively avoid costly turbulence events both from a financial as well as a brand erosion aspect. Significant turbulence events often make headlines, and when they do, the traveling public's fear level increases. The already serious turbulence forecast problem is expected to worsen as the climate warms (Aviation Turbulence, Williams and Joshi 2016). The industry needs good turbulence forecasts, better than those supplied to it today. A physically caused turbulence conceptual model such as the one presented in this report stands to be more successful than a statistically based one.

References

Byers, H.R. and R.B. Braham, 1949: *The Thunderstorm.* U.S. Government Printing Office, Washington DC. pp. 43.

Cornman, L. B., C. S. Morse, and G. Cunning, 1995: Real-time estimation of atmospheric turbulence severity from in-situ aircraft measurements. *J. Aircraft*, **32** (1), 171–177.

Doswell, C.A. III, R. Davies-Jones, and D.L. Keller: On summary measures of skill in rare event forecasting based on contingency tables, *Wea. Forecasting*, **5**, 576-585.

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

Lennartson, D. W. and D.W. McCann, 2014: Integrated turbulence forecasts, *Fourth Aviation, Range, and Aerospace Meteorology Special Symposium*, https://ams.confex.com/ams/94Annual/webprogram/Paper230413.html

McCann, D.W., 2001: A simple turbulence kinetic energy equation and aircraft boundary layer turbulence. *Natl. Wea. Digest*, **25**, 13-19.

McCann, D.W., 2006: Diagnosing and forecasting aircraft turbulence with steepening mountain waves. *Natl. Wea. Digest*, **30**, 77-92.

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

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

Sharman, R., and J. Pearson, 2017: Prediction of energy dissipation rates for aviation turbulence: Part I. Forecasting Non-convective turbulence. J. *Appl.Meteor. Climatol.*, **56**, 317-337.

Williams, J.K., G. E. Blackburn, J. A. Craig, R. K. Goodrich, J. Johnson, F. McDonough, G. Meymaris, J. M. Pearson, and R. D. Sharman, 2013: Identifying upper-level wake vortex encounters using routine turbulence reports. *Proc. 16th Conf. on Aviation, Range, and Aerospace Meteorology*, Austin TX, Amer. Meteor. Soc., Available at https://ams.confex.com/ams/93Annual/webprogram/Paper219378.html.

Williams, P.D., T.W.N. Haine, and P.L. Read, 2008: Inertia-gravity waves emitted from balanced flow: Observations, properties, and consequences. *J. Atmos. Sci*., **65**, 3543- 3556.

Williams. P.D. and M.M. Joshi, 2016: Clear-air turbulence in a changing climate. *Aviation Turbulence*, Springer International Publishing, Switzerland, ISBN 978-3-319-23629-2, Chapter 23, 465-480.

Wilson, E.N., 2012: *Case studies of clear-air turbulence: evaluation and verification of new forecasting techniques*. M.S. thesis, University of Georgia, 127 pp.