Multi-Model Ensemble Wake Vortex Prediction

Stephan Körner[*]

Deutsches Zentrum für Luft- und Raumfahrt, Oberpfaffenhofen, Germany

Nash'at N. Ahmad[†]

NASA Langley Research Center, Hampton, Virginia 23681

Frank Holzäpfel[‡]

Deutsches Zentrum für Luft- und Raumfahrt, Oberpfaffenhofen, Germany

Randal L. VanValkenburg[§]

NASA Langley Research Center, Hampton, Virginia 23681

Several multi-model ensemble methods are investigated for predicting wake vortex transport and decay. This study is a joint effort between National Aeronautics and Space Administration and Deutsches Zentrum für Luft- und Raumfahrt to develop a multi-model ensemble capability using their wake models. An overview of different multi-model ensemble methods and their feasibility for wake applications is presented. The methods include Reliability Ensemble Averaging, Bayesian Model Averaging, and Monte Carlo Simulations. The methodologies are evaluated using data from wake vortex field experiments.

Nomenclature

$\tilde{A}$	=	REA average
B_i	=	bias of model i
b	=	vortex spacing
b_0	=	initial vortex pair separation
D_i	=	absolute distance to ensemble mean of model i
$\tilde{\delta}_f$	=	ensemble forecast uncertainty
e	=	ensemble
ε	=	dimensional eddy dissipation rate
f_i	=	forecast of ith model
$\tilde{f}$	=	forecast average
f_+	=	upper ensemble uncertainty limit
f_-	=	lower ensemble uncertainty limit
g	=	gravitational acceleration
Γ	=	vortex circulation
Γ_0	=	initial vortex circulation
Γ^*	=	vortex circulation strength normalized by initial vortex strength, Γ_0
m	=	weighting factor for $R_{B,i}$
nv	=	natural variability
n	=	weighting factor for $R_{D,i}$

---

[*] Research Aerospace Engineer, DLR, Oberpfaffenhofen, Germany.
[†] Research Aerospace Engineer, NASA, Hampton, Virginia. Senior Member, AIAA.
[‡] Research Aerospace Engineer, DLR, Oberpfaffenhofen, Germany.
[§] Research Aerospace Engineer, NASA, Hampton, Virginia.

American Institute of Aeronautics and Astronautics

N	=	dimensional Brunt-Väisälä frequency
p	=	parameter
θ	=	potential temperature
$R_{D,i}$	=	weighting factor for model i (considering model convergence)
$R_{B,i}$	=	weighting factor for model i (considering model bias)
R_i	=	total weighting factor for model i
σ	=	standard deviation
σ_{obs}	=	variability calculated from measurements
σ_{err}	=	lidar measurement error
σ_{nv}	=	natural variability of vortices
t	=	vortex age
t_0	=	normalized vortex age
T	=	non-dimensional time
u	=	crosswind
V_0	=	initial vortex descent velocity
w	=	vortex descent speed
y	=	lateral position of the vortex core
$y*$	=	vortex lateral transport normalized by b_0
z	=	vertical position of the vortex core
$z*$	=	vortex height normalized by b_0
AGL	=	Above Ground Level
ANOVA	=	Analysis of Variance
APA	=	AVOSS Prediction Algorithm
ASOS	=	Automated Surface Observations System
ATM	=	Air Traffic Management
AVOSS	=	Aircraft VOrtex Spacing System
D2P	=	Deterministic 2-Phase Model
DEN03	=	Denver 2003 Wake Vortex Field Experiment (NASA)
DFW97	=	Dallas/Fort Worth 1997 Wake Vortex Field Experiment (NASA)
DLR	=	Deutsches Zentrum für Luft- und Raumfahrt
IGE	=	In-Ground Effect
LES	=	Large Eddy Simulation
Lidar	=	LIght Detection And Ranging
MEM95	=	Memphis 1995 Wake Vortex Field Experiment (NASA)
MCS	=	Monte-Carlo Simulation
MME	=	Multi-Model Ensemble
NASA	=	National Aeronautics and Space Administration
NGE	=	Near-Ground Effect
OGE	=	Out-of-Ground Effect
pdd	=	probability density distribution
P2P	=	Probabilistic 2-Phase Model
PL	=	Pulsed Lidar
PPE	=	Perturbed Physics Ensemble
RASS	=	Radio Acoustic Sounding System
REA	=	Reliability Ensemble Averaging
BMA	=	Bayesian Model Averaging
rmse	=	root mean square error
SME	=	Single Model Ensemble
TASS	=	Terminal Area Simulation System
TDAWP	=	TASS Driven Algorithms for Wake Prediction
TDP	=	Abbreviation used for the TDAWP model
WakeMUC	=	Munich Wake Vortex Field Experiment (DLR)
WakeFRA	=	Frankfurt Wake Vortex Field Experiment (DLR)
WakeOP	=	Oberpfaffenhofen Wake Vortex Field Experiment (DLR)

American Institute of Aeronautics and Astronautics

I. Introduction

STEADILY increasing air traffic is a challenge for air traffic controllers, airports, air carriers, and pilots. Safe operations may be threatened by wake vortices which can be encountered en-route, during departures and arrivals, and particularly in ground proximity, limiting the capacity of runways and posing a potential risk for aircraft. Atmospheric parameters such as wind conditions, thermal stratification, shear layers, and turbulence strongly affect wake vortex behavior. Dangerous situations can occur when wake vortices hover in the glide path corridor, drift to parallel runways, or live longer than expected. In order to predict and avoid these situations, wake vortex prediction models have been developed.

However, forecasts are associated with initial condition and model uncertainties, which results in the need for probabilistic forecasts. Multi-Model Ensemble (MME) approaches can cover model uncertainty and can be extended to also consider initial condition uncertainty. Several studies, many of them in the field of meteorology, have shown that combining the forecasts of structurally different models can improve both the deterministic and probabilistic forecast skill on average (Raftery et al. 2005). The success of the ensemble approach is based on the fact that any member model may be the best in certain situations (Hagedorn et al. 2005), so that together the models cover the full possible solution space. Certainly, employing the best member instead is tempting. However, the best member varies in different conditions and is therefore difficult to identify (Hagedorn et al. 2005).

In this paper we investigate whether ensemble methods can be used to increase the prediction skill of individual wake models by combining several predictions of wake decay and transport in an effective manner. We evaluate various distinct MME approaches, consisting of Reliability Ensemble Averaging, Bayesian Model Averaging, and Monte Carlo Simulation, which utilize the ensemble spread to predict probabilistic envelopes. Their prediction skills are evaluated using a test dataset and compared with the individual models' prediction skill. The models that are used were exchanged under an inter-agency NASA-DLR cooperation agreement. Wake measurement data from lidar (light detection and ranging) at Memphis (MEM95) (Campbell et al. 1997), Dallas (DFW97) (Proctor 98), Denver (DEN03) (Proctor 98), Frankfurt (WakeFRA) (Frech and Holzäpfel 2008), Munich (WakeMUC) (Holzäpfel et al. 2014) and at Oberpfaffenhofen (WakeOP) (Holzäpfel et al. 2014) are used for training and evaluation. The Memphis dataset contains wake history in all three phases of vortex descent (out-of-ground effect, near-ground effect, and in-ground effect) while the DLR data comprises vortices developing in ground effect. In addition to these data sets Dallas (DFW97) and Denver 2003 (DEN03) datasets were also used in model evaluation. All data are treated non-dimensionally: Γ is normalized by Γ_0, and transport and vertical descent by b_0.

II. Fast-Time Wake Transport and Decay Models

Fast-time wake models are empirical or semi-empirical algorithms used for real-time predictions of wake transport and decay based on aircraft parameters and ambient weather conditions. The aircraft dependent parameters include the initial circulation (Γ_0), the initial vortex descent velocity (V_0), and the vortex pair separation distance (b_0). The atmospheric initial conditions include vertical profiles of either temperature or potential temperature (θ), EDR (ε), and wind. The atmospheric parameters that affect wake decay are atmospheric stratification and turbulence. The models assume that all vorticity created due to lift is rolled up into a pair of counter-rotating vortices at the time of model initialization. The four models used in this study include: AVOSS (Aircraft Vortex Spacing System) Prediction Algorithm (APA), TASS (Terminal Area Simulation System) Derived Algorithms for Wake Prediction (TDP), Deterministic 2-Phased (D2P), and the Probabilistic 2-Phased (P2P). The APA and TDP models have been developed by the National Aeronautics and Space Administration (NASA), while the D2P and P2P models have been developed by the Deutsches Zentrum für Luft- und Raumfahrt (DLR). A brief overview of the four fast-time wake models is given in this section.

A. Deterministic 2-Phase (D2P) and Probabilistic 2-Phase (P2P) Wake Vortex Models

D2P and P2P wake vortex decay and transport models were developed at DLR (Holzäpfel 2003). Vortex decay in D2P is based on an analytical solution of the Navier-Stokes equations for the decaying potential vortex which has been extended to model the turbulent decay of a pair of counter-rotating vortices. The decay is divided into a diffusion and a rapid decay phase and has been adapted to the decay characteristics predicted by large eddy simulations. Moreover, it uses 5-15 m average circulation values in the vortex decay calculations. Vortex descent speed depends on a nonlinear relation to the circulation considering effects of stable thermal stratification. When the vortices approach the ground they start to slow down their descent and diverge laterally. This effect is modeled by introducing image vortices for the impenetrable ground boundary condition. With superposition of the induced velocity fields according to the law of Biot-Savart, the vortex divergence is reproduced. Secondary and tertiary vorticity that detaches

American Institute of Aeronautics and Astronautics

during this ground interaction is modeled using point vortices that rotate around the primary vortices together with their respective image vortices.

Crosswind promotes the detachment of secondary vorticity for the lee vortex and suppresses it for the luff vortex. The strength of the secondary vortices is adapted according to crosswind speed (Holzäpfel and Steen 2007). Measurements from WakeMUC showed that the lee vortex decays in average slightly faster than the luff vortex (Holzäpfel et al. 2014). This effect however is disregarded in the model physics. P2P is the probabilistic version of D2P with uncertainty allowances for the predicted parameters. Probability density functions (PDFs) derived from measurements relative to the forecasts are used to derive the level of probability for these limits. By shifting the uncertainty limits, theoretically any desired probability level can be attained. In this study we only employ the deterministic version D2P.

B. AVOSS (Aircraft Vortex Spacing System) Prediction Algorithm (APA)

In the late 1990s, under NASA's Aircraft Vortex Spacing System (AVOSS) project, significant advances were made in wake vortex modeling based on the data from field experiments and large eddy simulations. The initial versions of the AVOSS Wake Vortex Prediction Algorithm (APA) were developed during the AVOSS program and demonstrated in the AVOSS demo at the Dallas/Ft. Worth (Hinton 2001). The APA model computes the out-of-ground-effect (OGE) decay and descent based on Sarpkaya (Sarpkaya 2000; Sarpkaya et al. 2001). The model has an algorithm for enhanced rate of decay during the ground effect developed by Proctor et al. (2000). The scheme to compute lateral vortex transport is based on the vertical profile of crosswind (Robins and Delisi 2002), and the in-ground-effect (IGE) transport accounts for vortex spreading and rebound (Robins et al. 2002). The code development of APA is described in Robins and Delisi (2002).

C. TASS (Terminal Area Simulation System) Driven Algorithms for Wake Prediction (TDP)

The TDP model (Proctor et al. 2006; Proctor and Hamilton 2009) has been developed from parametric studies using large eddy simulation of wake vortices (Han et al. 2000). The TDP model uses separate prognostic equations based on the Terminal Area Simulation System (Proctor 1987) for vortex descent rate and circulation. The effect of crosswind shear on vortex descent rate is also taken into consideration, which allows the modeling of vortex tilt and change in lateral separation due to crosswind (Proctor 2009; Proctor and Hamilton, 2009). The IGE model in TDP is same as the one used in APA.

D. Comparison of Models

In order to show that the models fulfill the ensemble criterion of being structurally independent, a detailed model evaluation was conducted using data from four different field experiments (Appendix A). Model bias data based on an evaluation of all datasets were used to generate the overall statistics. Circulation data were normalized by Γ_0 and the lateral transport and vertical descent data were normalized by b_0. Distributions of data are shown in Figures 1-3 as a combined box-and-whisker plot with a mean-standard deviation overlay. The mean and standard deviation are shown by a green rectangle extended plus and minus one standard deviation from the central green line representing the mean of the distribution. The box portion shows the median (red vertical line) and the interquartile distance in both directions. The interquartile distances represent the 25-50 percentile and 50-75 percentile of the data, and are not necessarily equal in size. The notch in the box is an estimate of the significant differences. If the median of one distribution falls outside the notch on another, then the difference between the two distributions is statistically significant. On each side of the box the whiskers extend 1.5 times the interquartile distance for that side. Any data points falling outside the whiskers are shown individually as a red cross.

The statistics for the biases in circulation prediction, lateral transport prediction and vortex descent predictions are given in Tables 1-3. APA v3.2 had the smallest mean and D2P had the smallest standard deviation in the mean circulation bias. For lateral transport, D2P had the smallest mean and standard deviation in the bias. TDP v2.1 had the smallest mean and APA v3.2 had the smallest standard deviation in the vortex descent prediction bias.

For all four fast-time wake models, ANOVA tests were performed to identify any statistically significant differences in the model prediction bias. The results of ANOVA analysis are shown in Figures 1-3. Statistically significant differences were found in the bias (circulation, vortex descent) as a function of different models which are identified in the figures. APA v3.2 and v3.4 had statistically significant differences in vortex descent bias compared with D2P and TDP v2.1. Both D2P and TDP models are based on data derived from large eddy simulation, whereas the APA models are based primarily on observations. No statistically significant differences were identified in lateral transport predictions.

Table 1: Bias in Circulation Prediction (2734 points)

Option	mean	σ	median	2.5 percentile	97.5 percentile
APA32	0.007	0.226	0.042	-0.421	0.324
APA34	0.071	0.222	0.113	-0.375	0.366
TDP21	0.147	0.224	0.184	-0.300	0.428
D2P	0.111	0.217	0.148	-0.327	0.396
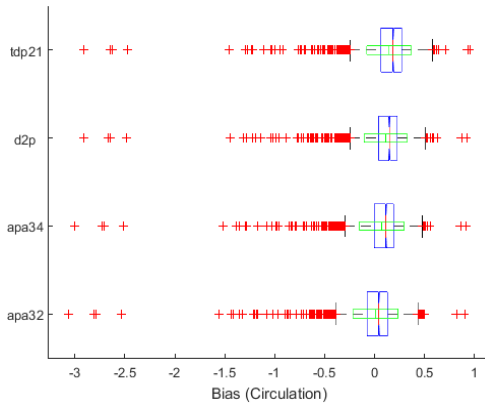
Figure 1. *Left panel shows box-and-whisker plot for bias in circulation prediction across all models. The right panel shows the statistically significant differences between models (Bias$_T$) based on the results of ANOVA analysis.*

Table 2: Bias in Prediction of Lateral Transport (2734 points)

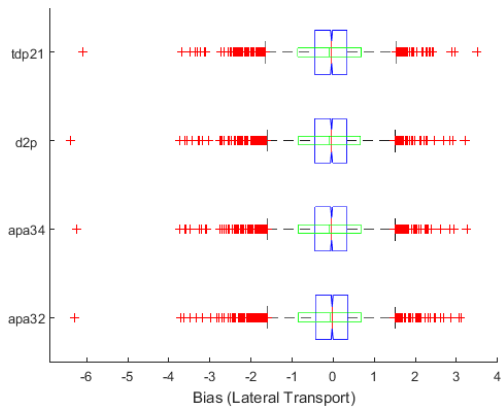| Option	mean	σ	median	2.5 percentile	97.5 percentile
APA32	-0.081	0.767	-0.026	-1.824	1.459
APA34	-0.088	0.758	-0.040	-1.796	1.399
TDP21	-0.091	0.771	-0.040	-1.822	1.446
D2P	-0.069	0.754	-0.035	-1.775	1.375

Figure 2. *Left panel shows box-and-whisker plot for bias in lateral transport prediction across all models. No statistically significant differences were found between models (Bias$_y$) based on the results of ANOVA analysis.*

5

Table 3: Bias in Prediction of Vortex Descent (2734 points)

Option	mean	σ	median	2.5 percentile	97.5 percentile
APA32	0.191	0.500	0.187	-0.888	1.173
APA34	0.160	0.504	0.166	-0.932	1.137
TDP21	0.068	0.502	0.096	-1.051	1.028
D2P	0.074	0.522	0.088	-1.023	1.051
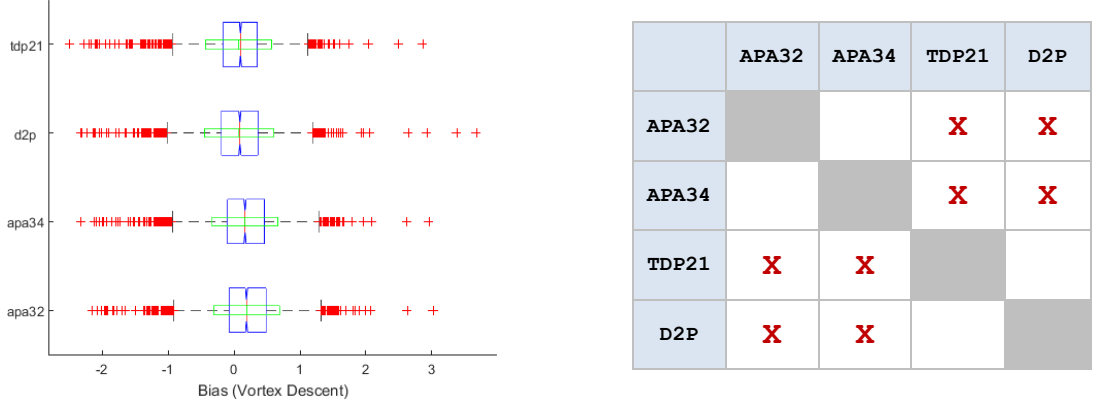
Figure 3. *Left panel shows box-and-whisker plot for bias in vortex descent prediction across all models. The right panel shows the statistically significant differences between models (Bias$_z$) based on the results of ANOVA analysis.*

III.   Multi-Model Ensemble Methods and Application

In this section several MME methods with both deterministic and probabilistic output are introduced and their applicability is evaluated. If necessary for the method, the ensemble is first trained with a set of 206 landings and then scored with a set of 200 independent landings. The whole dataset consists of selected high quality measurements with $z_0<1.7$ from the WakeMUC (Holzäpfel et al. 2014) and the WakeFRA (Frech and Holzäpfel 2008) campaigns. From this the training set and the test set are generated by randomly adding landings to each of them in order to cover similar ambient conditions, orography, and measurement uncertainties in both. The limit of 1.7 b_0 marks the boundary between Near-Ground-Effect (NGE) and Out-of-Ground-Effect (OGE). Ensemble performance in OGE has not yet been evaluated.

As stated above, a Multi-Model Ensemble needs structurally independent models. The ANOVA analysis, however, shows that the differences for the forecast of the vertical vortex position *z* between APA 3.4 and APA 3.2 are not statistically significant in IGE and NGE. As a consequence APA 3.2 is only considered for the *y*- and Γ-forecast when employing the Direct Ensemble Average (DEA), the Reliability Ensemble Average (REA), and the Bayesian Model Average (BMA). Moreover, no model exhibits statistically significant differences for the y-prediction, as lateral transport depends for the most part on the crosswind at the current vertical vortex position. Additionally, the wind exhibits strong variability which cannot be covered by the 10 minute averages used. Thus the models might be rated differently in terms of *y*-forecast for various training datasets, which is not necessarily related to their skill. Nevertheless the ensemble approach shall also be applied to the y-forecast, as a skillful probabilistic envelope might still deliver reasonable results. As skill metrics we apply the rmse and the bias. Furthermore we introduce the skill factor *s*, which takes the rmse of all parameters *p* into account (see equation (1)). Negative values of s_i indicate that model *i* is on average outperformed by the ensemble *e*.

$$s_i = \frac{\sum_{p=1}^{n} rmse_{e.p} / rmse_{i,p}}{n} - 1$$

(1)

American Institute of Aeronautics and Astronautics

A. Direct Ensemble Average (DEA)

In this method, the ensemble members are pooled together and their forecasts contribute equally weighted to the MME forecast (see equation (1)). However, this approach neglects the differences in quality of the individual ensemble members' performances. Nevertheless, it is used as a baseline with which to compare the outputs of the REA approach.

$$\bar{f} = \frac{\sum_i f_i}{i} \tag{2}$$

Table 4 shows that the DEA approach does not improve the forecast of any parameter in comparison to the respective best member. The skill factors support this finding.

Table 4: performance of the DEA ensemble and its individual members (median for 200 test cases). Skill $s_{z,y,\Gamma}$ considers all parameters, skill $s_{z,\Gamma}$ only the z and Γ forecast.

	rmse Γ_{luff}	rmse Γ_{lee}	rmse y_{luff}	rmse y_{lee}	rmse z_{luff}	rmse z_{lee}	skill $s_{z,y,\Gamma}$	skill $s_{z,\Gamma}$
DEA	0.115	0.104	0.829	0.566	0.199	0.167	0.00	0.00
TDP 2.1	0.099	0.096	0.901	0.622	0.257	0.176	-0.033	-0.008
APA 3.4	0.149	0.128	0.949	0.561	0.212	0.185	-0.115	-0.144
APA 3.2	0.218	0.180	0.989	0.548	(0.201)	(0.188)	-0.191	-0.254
D2P	0.103	0.106	0.571	0.583	0.136	0.162	0.168	0.146

B. Reliability Ensemble Averaging (REA)

Giorgi and Mearns (2002) have developed an ensemble method, which rates the individual forecasts according to their bias in a set of training data. Additionally, they assume in their approach that the distance of each forecast to the ensemble mean is a measure of its reliability. They found that this approach allows a reduction of the uncertainty range as the influence of poorly performing models can be decreased. Instead of applying this to the average change of temperature ΔT as in Giorgi and Mearns (2002) we aim for the average $\tilde{f}$ of the individual forecasts f_i of each model i:

$$\tilde{f} = \frac{\sum_i R_i f_i}{\sum_i R_i} \tag{3}$$

where R_i denotes the general reliability factor for model i and is used to weight its forecast. The general reliability factor R_i consists of the two specific factors $R_{B,i}$ and $R_{D,i}$ that are here determined by

$$R_i = \left[\left(R_{B,i} \right)^m \cdot \left(R_{D,i} \right)^n \right]^{1/(m \cdot n)} = \left\{ \left[\frac{\min(B_i)}{|B_i|} \right]^m \left[\frac{nv}{|D_i|} \right]^n \right\}^{[1/(m \cdot n)]} \tag{4}$$

$R_{B,i}$ (performance criterion) is a factor that describes the model reliability as a function of model bias B_i, and $R_{D,i}$ (convergence criterion) characterizes the model reliability in terms of distance between an individual model forecast and the ensemble prediction D_i. The performance and convergence criterion are determined individually for decay, descent, and lateral transport of the luff and lee vortices, respectively. nv denotes a measure of natural variability and ensures that models with a deviation less than nv are regarded as reliable, so that $R_{D,i} = 1 = R_i$. In other words, if the bias or the absolute distance is smaller than nv, the corresponding reliability factor is set to 1, preventing R_i from exceeding unity. In the original approach B_i is also calculated using the natural variability. However in our application, it is only applicable to D_i in order to prevent it from becoming too large. Applying it to B_i impairs the ensemble forecast.

The distance $D_{T,i}$ is calculated iteratively for every single forecast with the direct ensemble average as first guess. Both the performance criterion and the convergence criterion can be weighted by the exponents n and m. In order to estimate the uncertainty of the ensemble forecast, Giorgi and Mearns (2002) introduce $\tilde{\delta}_f$, which quantifies how

strong the individual models deviate from the ensemble forecast, considering their respective weights (see equation (5)).

$$\tilde{\delta}_f = \left[\tilde{A}\left(f_i - \tilde{f}\right)^2 \right]^{1/2} = \left[\frac{\sum_i R_i \left(f_i - \tilde{f}\right)^2}{\sum_i R_i} \right]^{1/2} \tag{5}$$

Applied against the ensemble mean we obtain an upper and lower uncertainty limit.

$$f_\pm = \tilde{f} \pm \tilde{\delta}_f \tag{6}$$

Furthermore the collective reliability $\tilde{\rho}$ of the ensemble forecast is given by the weighted average of the individual model reliability factors:

$$\tilde{\rho} = \frac{\sum_i R_i^2}{\sum_i R_i} \tag{7}$$

To summarize: we train the ensemble with the help of the model bias of a training dataset and additionally rate it during the operational run by a weight that depends on the individual model distance to the ensemble mean. A model is regarded as reliable, when it performs well in the training, if it is close to the ensemble mean and if it is forecasted within the natural variability.

1. Application

The REA method can only deliver fair results, if the reliability factors of the models are not all equal. In the case that the reliability factors all equal 1, because B_i or D_i are within the natural variability, the ensemble cannot be better than the DEA approach. By using the natural variability as a tuning parameter, we can achieve a uniform distribution of possible reliability factors from 0 to 1, without weighting the models equally. Other than for $R_{B,i}$ the natural variability cannot be neglected for $R_{D,i}$. As the ensemble mean may converge towards an individual model prediction, the respective $R_{D,i}$ must be prevented from becoming infinite. We obtain good results for $nv_{z*} = nv_{y*} = 0.06$ and $nv_{\Gamma*} = 0.04$, which is on the order of the model bias in the training set. Contrarily, $R_{B,i}$ is calculated in relation to the bias of the best model in every single simulation, which promotes skilled models stronger than if relating them to nv. To prove that both $R_{B,i}$ and $R_{D,i}$ are independent, they are plotted against each other. Figure 4 indicates that there is no correlation. This, and the fact that the presence of $R_{D,i}$ improves the MME skill, justify that the convergence and the performance criterion are necessary. For the test dataset the best results are obtained when both factors are equally weighted ($n = m = 1$).

Concerning the ambient weather dependency of $R_{B,i}$, the best skill was achieved with a total wind dependency for Γ and a crosswind dependency for z. This is a step into the direction of a best member approach, where for each situation a best model is identified. Regarding the forecast of the lateral position y it is difficult to obtain a rating of the y-forecast that is not only valid for the training dataset. This can be related to the crosswind variability that determines the quality of the prediction of lateral transport. However, a model which predicts the vertical vortex position correctly also has better wind information in the respective altitude. As a consequence the ensemble prediction of the lateral vortex position y is best when we assume that $R_{B,i,y} = R_{B,i,z}$.

The forecast reliability ρ is calculated for all parameters according to equation (7). If this measure was able to predict the ensemble forecast quality, a correlation between rmse and ρ should be visible. However, Figure 5 only shows a weak correlation regarding the *z*-prediction and no correlation for the *y*- and *Γ*-forecast. Consequently ρ is not applicable to quantify the reliability of wake vortex predictions, at least not in the way it is formulated in equation (7).
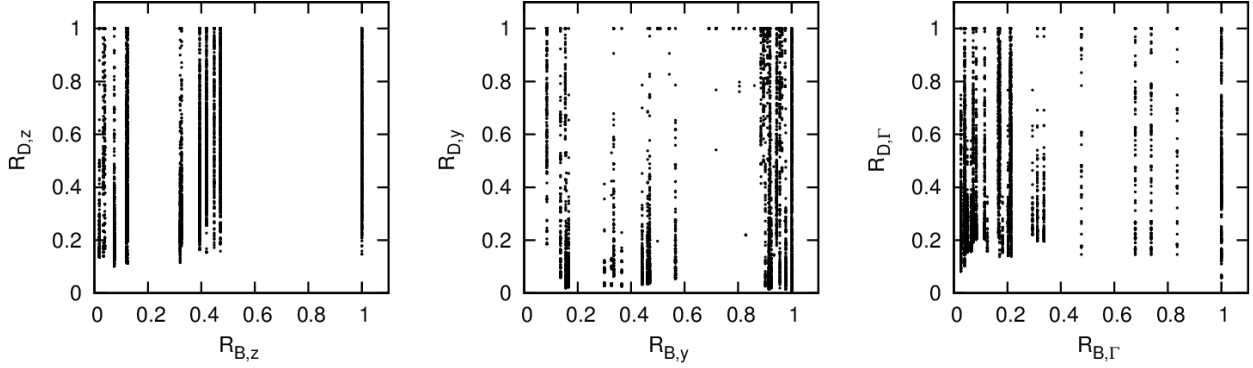
American Institute of Aeronautics and Astronautics

Figure 4. R_D **against** R_B **for all models and for the test dataset.** *R_B and R_D are uncorrelated for all parameters indicating that both factors are valuable.*
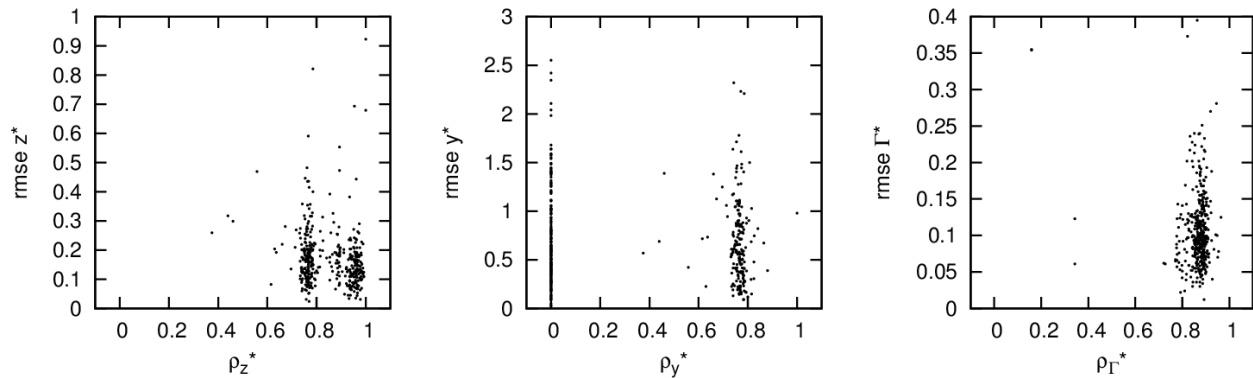

Figure 5. Correlation between the forecast reliability ρ and the ensemble rmse of the respective forecast. *A weak correlation can be found for z, but no correlation exists for y and Γ.*

The uncertainty limits are first calculated as suggested by Giorgi and Mearns (2002) for luff and lee vortices separately. As upper and lower limit we use the conservative limit of either of them. However, this approach does only consider model uncertainty and not the uncertainty of the initial conditions. Hence, the initial condition uncertainty is taken into account by adding a priori determined standard deviations to the REA uncertainty limits. For WakeMUC, the initial conditions are derived from lidar measurements, so the uncertainties are the lidar measurement uncertainties $\sigma_{z0} = 9$ m, $\sigma_{y0} = 13$ m, and $\sigma_{\Gamma 0} = 13$ m²/s (Köpp et al. 2005). If not derived from lidar, the uncertainty for Γ_0 is calculated using the error propagation method considering uncertainties of aircraft mass, air density, and flight speed (σ_z=3 m and σ_y=7 m (Holzäpfel 2014), σ_m=1300 kg, σ_b=1.5 m, σ_ρ=0.0048 kg/m³, σ_v=4 m/s). Although in ground proximity the lower limits for vertical vortex position and for circulation are of no interest for operational use, they are computed here.

Figure 6 shows an ensemble forecast according to the REA method. Both the individual model predictions and the ensemble mean are depicted. If we look at the uncertainty envelopes, of yet unknown probability level, we see that they widen with ongoing time as expected. Nevertheless, the envelopes do not sufficiently represent the variability of the atmosphere. Consequently, the luff vortex that hovers over the runway, due to weaker crosswind than captured with the 10 minute wind average, does not lie within the uncertainty bounds of the lateral prediction. Figure 7 shows the reliability factors for this landing under the prevailing wind conditions.

From Table 5 we see that the ensemble is superior for Γ_{luff} and second best for Γ_{lee}, y_{luff}, y_{lee}, z_{luff} and z_{lee}. Translated into a skill factor, the ensemble outperforms the best member by 1.4 % if we include the *y*-forecast. Otherwise a skill improvement of 3.1 % can be achieved.
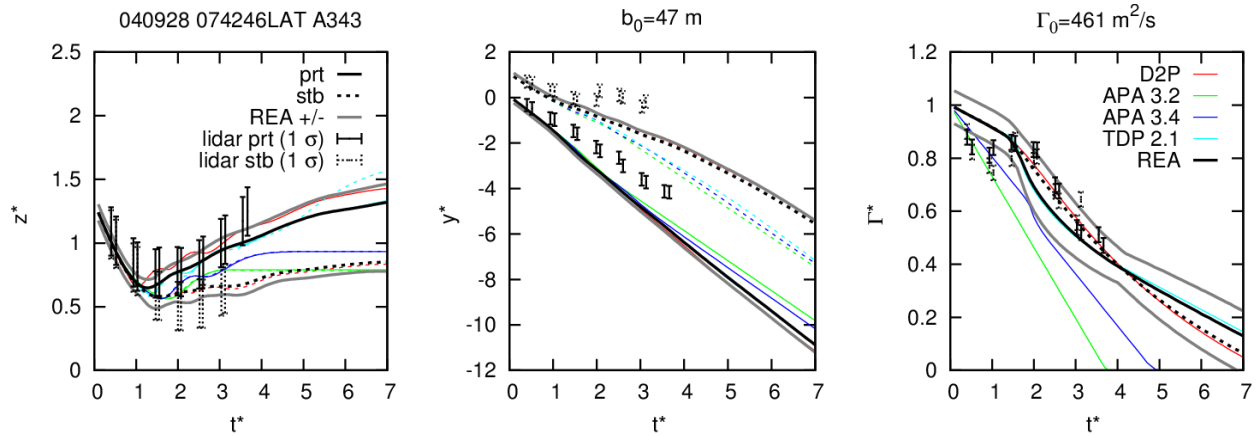
American Institute of Aeronautics and Astronautics

Figure 6. REA forecast for a single landing. *Lidar measurements are depicted as error bars that show the measurement uncertainty.*
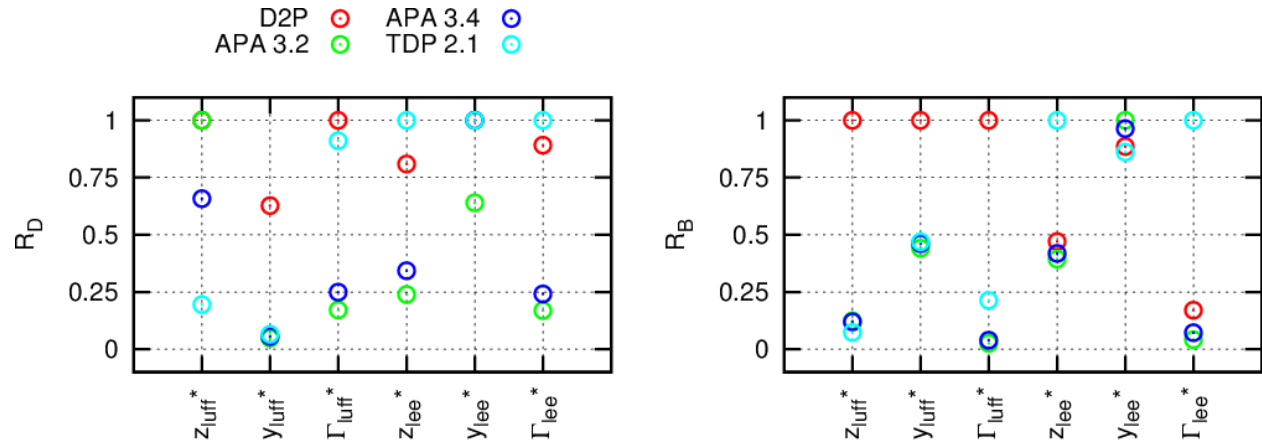

Figure 7. Reliability factors R_D and R_B for the individual members. *R_B is a global value that changes only for different ambient conditions. R_D, however, differs from simulation to simulation. This diagram corresponds to the landing shown in Figure 6. The green values of APA3.2 regarding z are only theoretical factors, as APA3.2 is not employed for the ensemble mean of the vertical position.*

Table 5: performance of the REA ensemble and its individual members (median for 200 test cases). Skill $s_{z,y,\Gamma}$ considers all parameters, skill $s_{z,\Gamma}$ only the z and Γ forecast.

	rmse Γ_{luff}	rmse Γ_{lee}	rmse y_{luff}	rmse y_{lee}	rmse z_{luff}	rmse z_{lee}	skill $s_{z,y,\Gamma}$	skill $s_{z,\Gamma}$
REA	0.095	0.097	0.576	0.589	0.137	0.169	0.00	0.00
TDP 2.1	0.099	0.096	0.901	0.622	0.257	0.176	-0.159	-0.136
APA 3.4	0.149	0.128	0.949	0.561	0.212	0.185	-0.232	-0.262
APA 3.2	0.218	0.180	0.989	0.548	0.201	0.188	-0.298	-0.362
D2P	0.103	0.106	0.571	0.583	0.136	0.162	-0.017	-0.031

American Institute of Aeronautics and Astronautics

C. Bayesian Model Averaging

The Bayesian Model Averaging (Hoeting et al. 1999) has been employed to temperature forecasts by Raftery et al. (2005). Instead of overall mean skill metrics, it employs the probability that a model is best, and its skill when it is the best among the ensemble. This can be formulated with the law of total probability. It describes the probability that B occurs in terms of the probability that A_n occurs, and the probability that B occurs given A_n (see equation (8)).

$$P(B) = \sum_n P(B \cap A_n) = \sum_n P(A_n)P(B \mid A_n) \tag{8}$$

Raftery et al. (2005) apply this to an ensemble forecast. In that example y denotes the quantity to be forecast and y^T denotes the training data for I different statistical models $M_1,...,M_I$. The forecast pdd is then given by

$$p(y) = \sum_{i=1,I} p(y \mid M_i)p(M_i y^T). \tag{9}$$

$p(y \mid M_i)$ is the forecast pdd of M_i alone and $p(M_i y^T)$ denotes the posterior probability of model M_i being correct given the training data. The resulting pdd $p(y)$ is then the sum of the individual model pdds, weighted by their posterior probabilities. Referring this to dynamic models, the pdd quantifies the uncertainty about the best member in our ensemble. Each of the forecasts $f_1,....,f_I$ by model i has a posterior probability w_i of being the best member among the ensemble and a conditional pdd $g_i(y \mid f_i)$. The latter describes the accuracy of f_i, given that it is the best forecast in the ensemble and is centered at the individual deterministic model forecast. The individual w_i's sum up to 1. The ensemble pdd is then defined by

$$p(y \mid f_1,....,f_I) = \sum_{i=1,I} w_i g_i(y \mid f_i). \tag{10}$$

1. Application

Raftery et al. (2005) apply a maximum likelihood method to optimize w_i and g_i on the basis of the training dataset. However, we proceed differently and compute w_i from the number of times that model i is best among the ensemble. Furthermore, the pdd g_i associated standard deviation σ_i of each model i is replaced by the corresponding model rmse for each parameter, assuming that g_i is normally distributed.

Both σ_i and w_i are calculated on the basis of the training dataset separately for each parameter and for both luff and lee vortices. Figure 8 shows the weighted pdds of all ensemble members at $t^*=0$ when all models are initialized. These pdds need to be computed for any time step ($\Delta t^*=0.1$) (see Figure 9) in order to derive the ensemble forecast (see Figure 10), where the pdd's mean value is given by the individual deterministic model prediction and its standard deviation by the model performance within the training dataset. Unlike the REA approach, here the model weights do not depend on the ambient conditions. Figure 9 shows, that the Γ-forecast is well-dispersed, in contrast to the prediction of z and y. According to the scoring on the basis of the test dataset in Table 6, the ensemble is superior for Γ_{luff} and Γ_{lee}, second best for y_{lee}, z_{luff} and z_{lee}, and second worst for y_{luff}. This leads to an improvement of skill by 1.8 % regarding all parameters, with respect to the best performing individual model. If only the z- and Γ-forecast are included, the ensemble even improves the skill by 4.3 %. So even though we do not include a best member approach by rating the models for different weather conditions, the BMA method delivers good results.
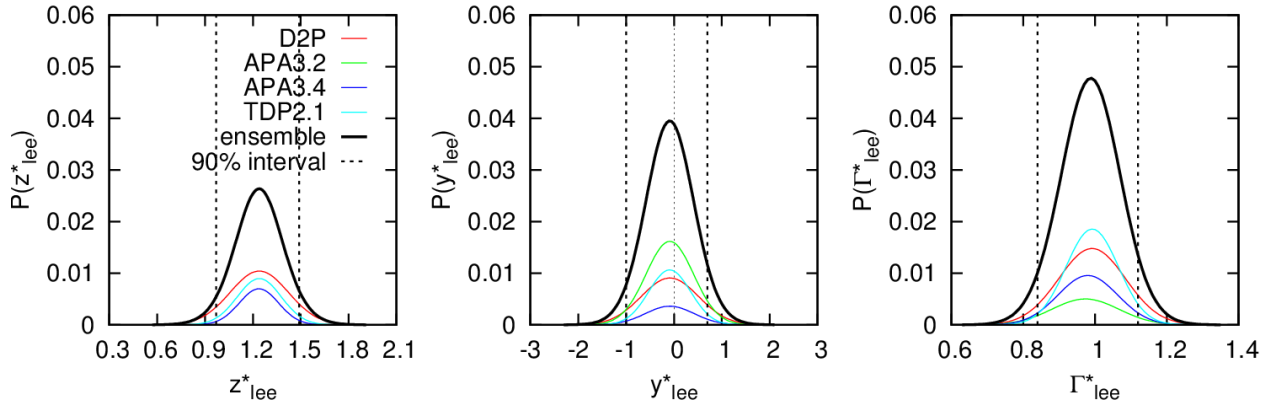
Figure 8. BMA pdd of the lee vortex forecast for t*=0. *All pdds are centered at the same value, which is for t*=0 equal to the initial conditions. The model pdds(colored lines) are weighted and then summed up to obtain the ensemble forecast (black curve). The black dotted lines denote the 90 % uncertainty interval.*
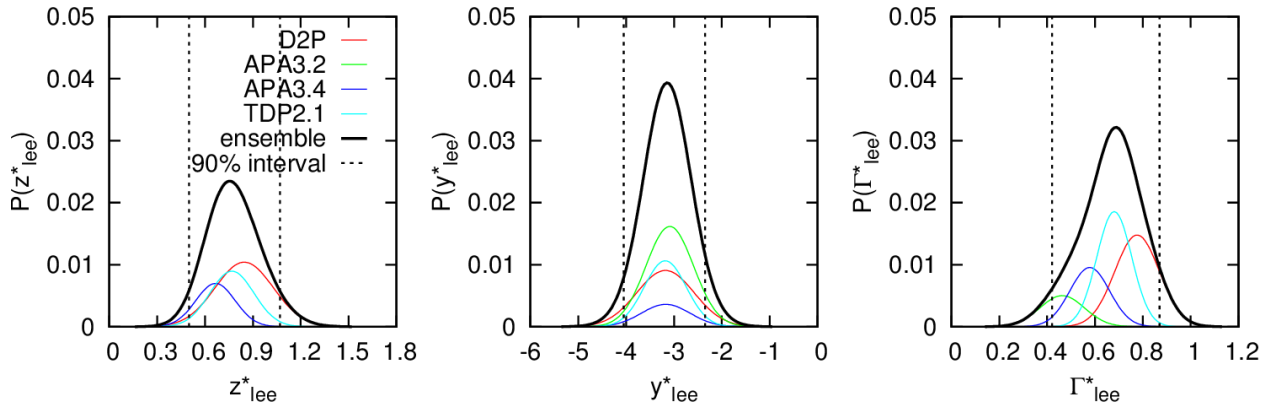

Figure 9. BMA pdd of the lee vortex forecast for t*=2. *The pdd centers are now shifted, as the models predict different values. As for t*=0, the individual pdds are weighted and then summed up. The black dotted lines denote the 90 % uncertainty interval.*
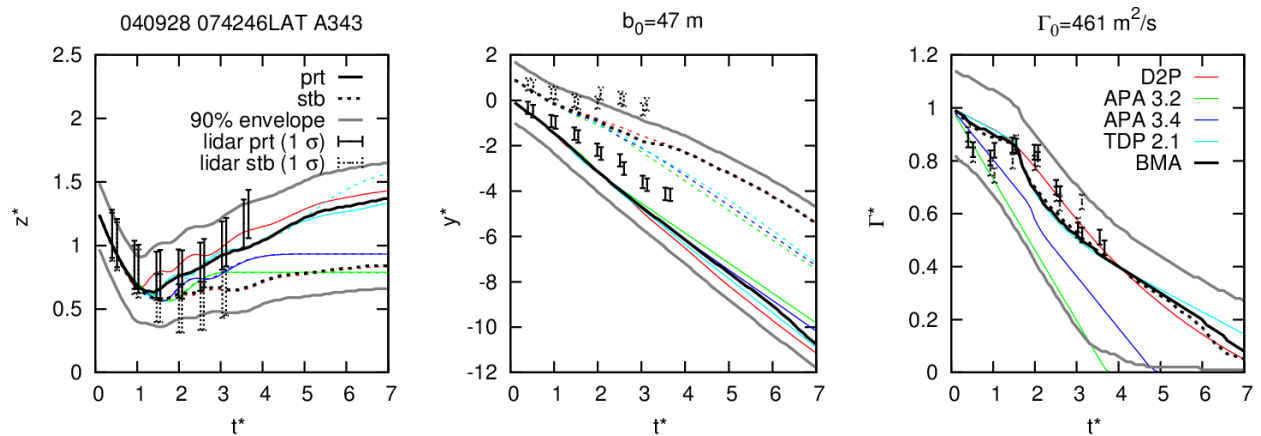

Figure 10.   BMA wake vortex forecast for a single landing. *No initial condition uncertainty was added to the uncertainty envelopes. By calculating σ from a training dataset that also exhibits uncertainty about the starting values, initial condition uncertainty is already considered beforehand.*

Table 6: Performance of the BMA ensemble and its individual members (median for 200 test cases). Skill $s_{z,y,\Gamma}$ considers all parameters, skill $s_{z,\Gamma}$ only the z and Γ forecast.

	rmse Γ_{luff}	rmse Γ_{lee}	rmse y_{luff}	rmse y_{lee}	rmse z_{luff}	rmse z_{lee}	skill $s_{z,y,\Gamma}$	skill $s_{z,\Gamma}$
BMA	0.092	0.093	0.626	0.565	0.139	0.168	0.00	0.00
TDP 2.1	0.099	0.096	0.901	0.622	0.257	0.176	-0.167	-0.151
APA 3.4	0.149	0.128	0.949	0.561	0.212	0.185	-0.237	-0.272
APA 3.2	0.218	0.180	0.989	0.548	0.201	0.188	-0.302	-0.368
D2P	0.103	0.106	0.571	0.583	0.136	0.162	-0.018	-0.043

D. Monte-Carlo Simulations

The previous sections introduced methods that cover model uncertainty. The additional uncertainty in the initial conditions is taken into account by either adding measurement errors to the uncertainty bounds, or it is considered in the training phase. In the case for the BMA method, the uncertainties in the initial conditions are included in the pdds. However, these approaches do not consider how the perturbed forecasts might develop in time due to model non-linearity. To study this effect Monte Carlo simulations are utilized. The results can then be used to compute probabilistic envelopes. Monte-Carlo method is simple to implement and can effectively account for uncertainties in the initial conditions. Computationally the method is expensive and although useful for systems level analysis of wake considerations in ATM concepts development, it may not be feasible for operational implementations.

2. Application

The sources of uncertainties in model initial conditions include aircraft dependent parameters (V_0 and b_0); initial vortex location (z_0 and y_0); and the ambient environment characterized by eddy dissipation rate (ε), stratification (N), and the crosswinds (u). If the probability density functions can be constructed, then they are utilized, otherwise uniform distributions are used within prescribed bounds.

In the current implementation, b_0 can either be held constant based on elliptical wing loading or varying between $0.95b_0$ and b_0 (Holzäpfel 2014). The initial circulation value is varied between 0.9-$1.2\Gamma_0$. Perturbations for y_0 are generated using a normal distribution with mean set to y_0 and σ=82ft (25m) (Holzäpfel 2014). Similarly for z_0, a normal distribution is used with mean set to z_0 and σ=23ft (7m). If the aircraft is in IGE, then σ is set to 13ft (4m).

The perturbations in the environmental initial conditions can be based on uniform distributions within prescribed bounds or obtained from probability density functions. An average of the vertical profile truncated at z_0 (height of vortex generation) is used as the mean. Once the averages of eddy dissipation rate and crosswinds have been calculated, k perturbations are generated using the probability density functions. Uniform weather profiles are generated from these perturbations. Simulations are performed with all inputs, and the standard deviation and the mean are calculated for each time step. The standard deviation is then added and subtracted to the mean to create bounds for circulation strength and vortex location.

An example from the Memphis 1995 wake vortex field experiment (Campbell 1997) is given in Figures 11-13 to illustrate the methodology. The perturbations in the initial location were introduced as described above. The initial circulation strength was held constant and the initial vortex spacing was assumed to be varying between $0.95b_0$ and b_0. The weather probability density functions were obtained from the Memphis 1995 weather data. The environmental initial conditions for Case 1995-08-10-233255 are given in Figure 11. The horizontal line in the Figure 11 plots indicate the height of vortex initialization. The mean was calculated by averaging the profile values below this height. The perturbations in initial conditions (Figure 12 – top row) were generated using the mean for the initial profile and the weather probability density functions. The probability density functions for this case are shown in Figure 12 (bottom row). The generation of the theta profiles is done from the potential temperature gradient probability density function.

A total of 400 random perturbations in initial conditions were generated and the results of the simulations are given in Figure 13 (top row). The mean, $\pm 2\sigma$ bounds, the deterministic APA v3.4 solution, and observations are shown in Figure 13 (bottom row). In this example only APA v3.4 was used. The prediction of vortex location is well bounded within the 2σ bounds. In the case of circulation prediction, several observation are outside the predicted 2σ bounds. However it should be noted that most of these circulation values are greater than the value of Γ_0 that was used to initialize the model.
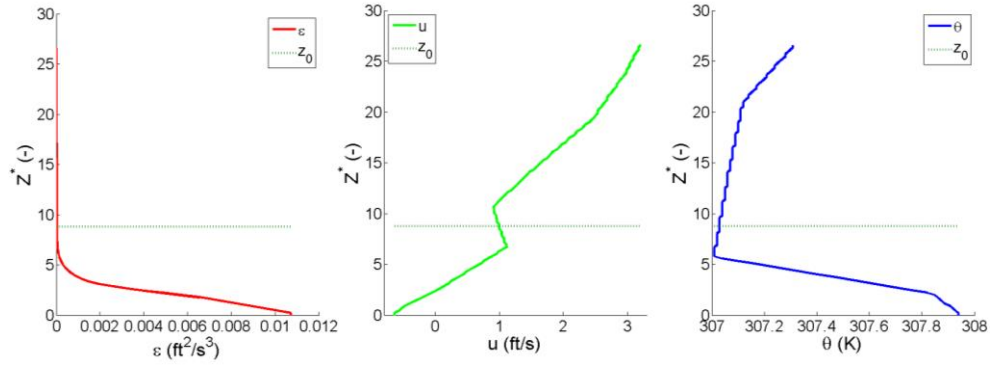
American Institute of Aeronautics and Astronautics

Figure 11.   **Case 1995-08-10-233255.** *EDR, crosswind and potential temperature initial conditions.*
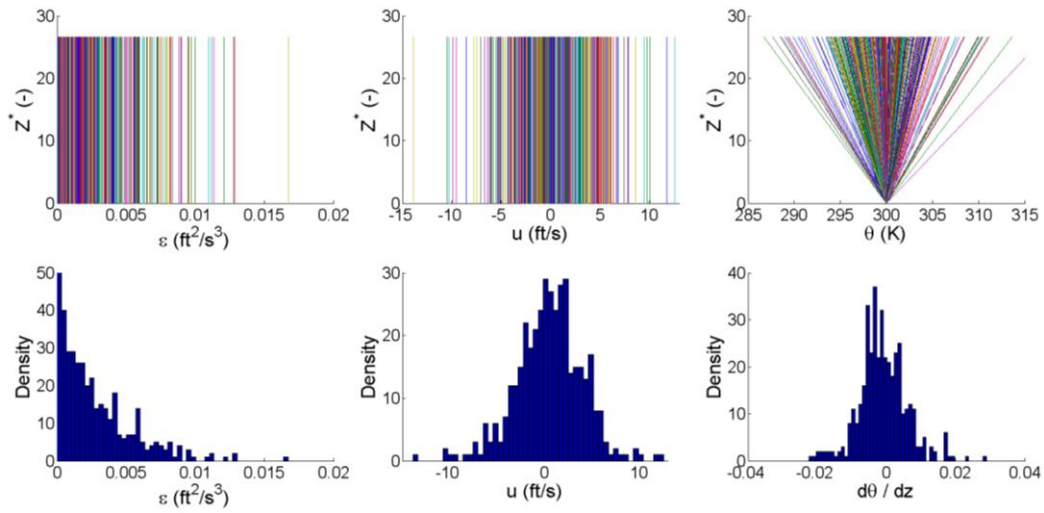

Figure12. Case 1995-08-10-233255. *Generated input profiles for EDR, crosswind and theta and the corresponding probability density functions.*
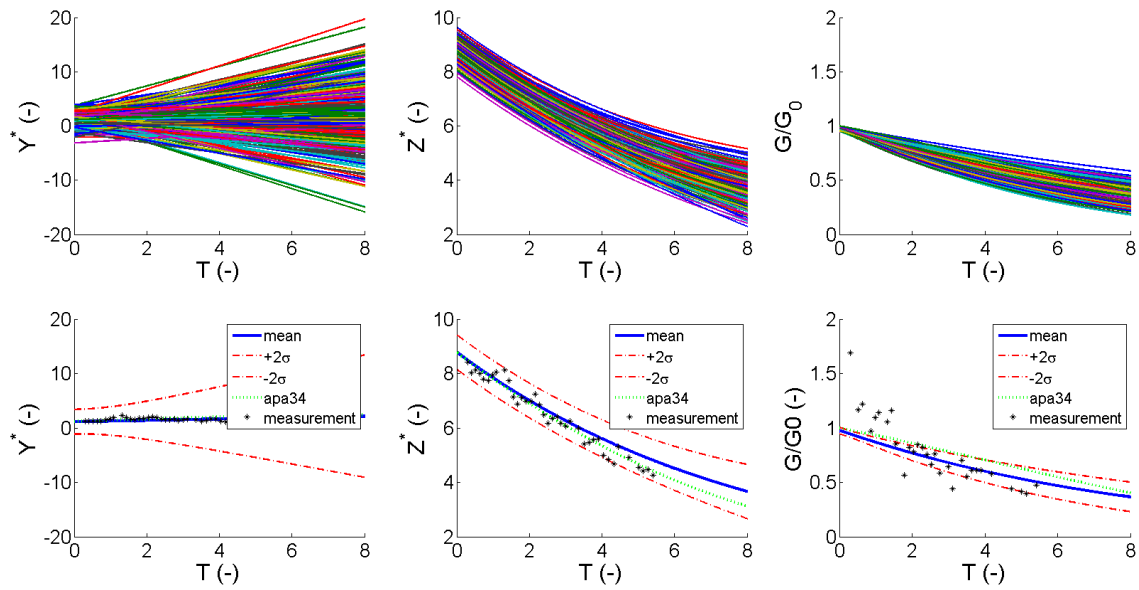

Figure 13.   **Case 1995-08-10-233255.** *The top row shows the results from all simulations. The bottom row shows the mean, and the bounds generated from the Monte-Carlo run.*

American Institute of Aeronautics and Astronautics

A sensitivity analysis was performed by looking at the convergence of the normalized root mean squared error (NRMSE) for different numbers of perturbations, k. The result given by using $k = 4096$ was used as the reference. Figure 14 shows the convergence of Γ, y, and z predictions. The error in Monte-Carlo prediction decreases by increasing the number of perturbations. The prediction of lateral transport is most sensitive to the number of perturbations and stays larger than Γ and z when k is increased.
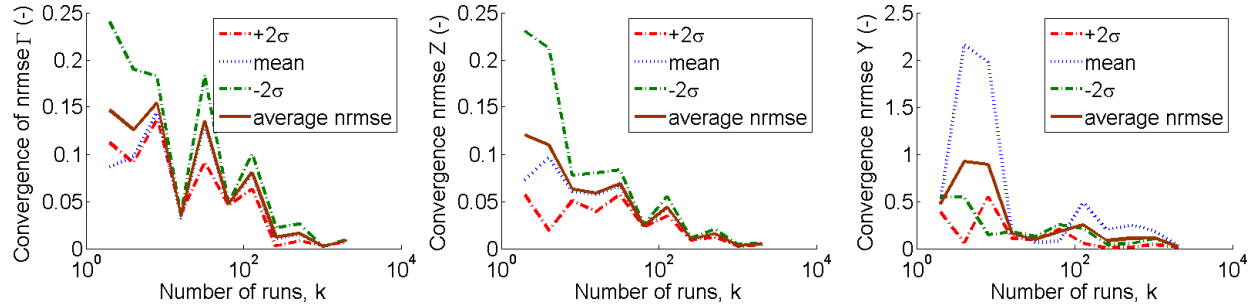

Figure 14. **Convergence of the NRMSE with increased number of simulations.** *Vortex circulation, Γ (left), vortex descent, z (middle), and Lateral transport, y (right).*

The Monte Carlo method was applied to four different wake vortex datasets: Memphis 1995 (MEM95), Dallas/Ft. Worth 1997 (DFW97), Denver 2003 (DEN2003), and WakeOP, to test its performance. Every case in the dataset that had lidar measurements for both vortices of the wake vortex pair was used to evaluate the method. This slightly reduced the total number of cases (wake tracks) which could be used in the evaluation. To determine the success rate of the Monte Carlo simulations, the number of measured vortices within the $\pm 2\sigma$ bounds of the predictions were counted and divided by the total number of vortices available in the case. In case of Γ, an additional metric was defined which looked at the number of measurements bounded by the maximum predicted circulation strength. For circulation strength this is a more meaningful metric compared to the $\pm 2\sigma$ bounds.

The four models were first ran individually with 10 perturbed initial conditions and then a multi-model ensemble was calculated from these four runs. Identical sets of perturbations were provided to all four models in these runs. The final time of simulation was determined by the shortest final simulation time amongst the four models. All other model outputs were truncated to match the shortest simulation time. This was required in order to construct the multi-model ensemble.

The success rates for all models and the multi-model ensemble are given in Tables 7-10 for different field experiments. In nearly all metrics and cases, the multi-model ensemble gave the highest success rates compared to individual fast-time models. In some cases the circulation prediction success rate of an individual model was significantly lower than the multi-model ensemble. The y-location was predicted with sufficient confidence with a success rate in the range of 90% or greater. Compared to the lateral transport, the success rate for vortex descent prediction was lower. For example using the MEM95 data, the success rates for vortex height was 63%, 64%, 69%, and 67% for APA v3.2, APA v3.4, TDP v2.1, and D2P respectively. As expected, the prediction of maximum bounded circulation was better compared to $\pm 2\sigma$ bounds prediction of circulation in all cases for all models and the multi-model ensemble. Another set of simulations (Run 2) was conducted in which a separate set of random initial conditions were provided for each model. The multi-model ensemble was then calculated from these simulations. This was done only using MEM95 data and the results are given in Table 11. The slight improvement in the multi-model ensemble predictions can be attributed to the increased number of initial perturbations (40 instead of 10).

Figures 15 and 16 show two Monte-Carlo simulations from the MEM95 set. A total of 10 random perturbations in initial conditions were generated for each model and are shown in the top row of Figures 15-16. The mean, $\pm 2\sigma$ bounds, the deterministic model solutions, and observations are shown in the bottom row of Figures 15-16. The prediction of vortex location is well bounded within the 2σ bounds. As in the previous example, several observation are outside the predicted 2σ bounds for circulation. Again it should be noted that these circulation values are greater than the value of Γ_0 that was used to initialize the model and their inclusion in model evaluation adversely effects the success rates as well as other metrics such as root mean square error and bias (Appendix A). The prediction of 2σ bounds for lateral transport was generally overly conservative in these simulations and can be improved by further refining the initial crosswind pdd.

American Institute of Aeronautics and Astronautics

Table 7: Success rates for models compared with the multi-model ensemble (MEM95)

Parameter	APA v3.2	APA v3.4	TDP v2.1	D2P	Multi Model
y	0.88	0.88	0.88	0.89	0.89
z	0.63	0.64	0.69	0.67	0.68
$\Gamma_{within\ bounds}$	0.48	0.50	0.53	0.48	0.66
$\Gamma_{under\ max}$	0.56	0.69	0.79	0.75	0.77

Table 8: Success rates for models compared with the multi-model ensemble (DFW97)

Parameter	APA v3.2	APA v3.4	TDP v2.1	D2P	Multi Model
y	0.99	0.99	0.99	0.99	0.99
z	0.72	0.71	0.72	0.73	0.73
$\Gamma_{within\ bounds}$	0.53	0.51	0.51	0.45	0.60
$\Gamma_{under\ max}$	0.66	0.73	0.81	0.78	0.78

Table 9: Success rates for models compared with the multi-model ensemble (DEN03)

Parameter	APA v3.2	APA v3.4	TDP v2.1	D2P	Multi Model
y	0.99	0.99	0.99	0.99	0.99
z	0.70	0.68	0.74	0.73	0.74
$\Gamma_{within\ bounds}$	0.56	0.48	0.35	0.33	0.58
$\Gamma_{under\ max}$	0.90	0.95	0.94	0.97	0.98

Table 10: Success rates for models compared with the multi-model ensemble (WakeOP)

Parameter	APA v3.2	APA v3.4	TDP v2.1	D2P	Multi Model
y	1.00	1.00	1.00	1.00	1.00
z	0.66	0.56	0.40	0.48	0.65
$\Gamma_{within\ bounds}$	0.60	0.64	0.71	0.52	0.98
$\Gamma_{under\ max}$	0.60	0.64	0.79	0.73	0.98

Table 11: Success rates for models compared with the multi-model ensemble (MEM95) – Run 2

Parameter	APA v3.2	APA v3.4	TDP v2.1	D2P	Multi Model
y	0.88	0.90	0.90	0.90	0.90
z	0.63	0.64	0.67	0.68	0.70
$\Gamma_{within\ bounds}$	0.48	0.48	0.54	0.48	0.67
$\Gamma_{under\ max}$	0.56	0.68	0.80	0.76	0.80

American Institute of Aeronautics and Astronautics

Figure 15.   **Case 1995-08-07-001041.** *The top row shows the results from all simulations. The bottom row shows the mean, and the bounds generated from the Monte-Carlo run along with the results of deterministic simulations of each model.*
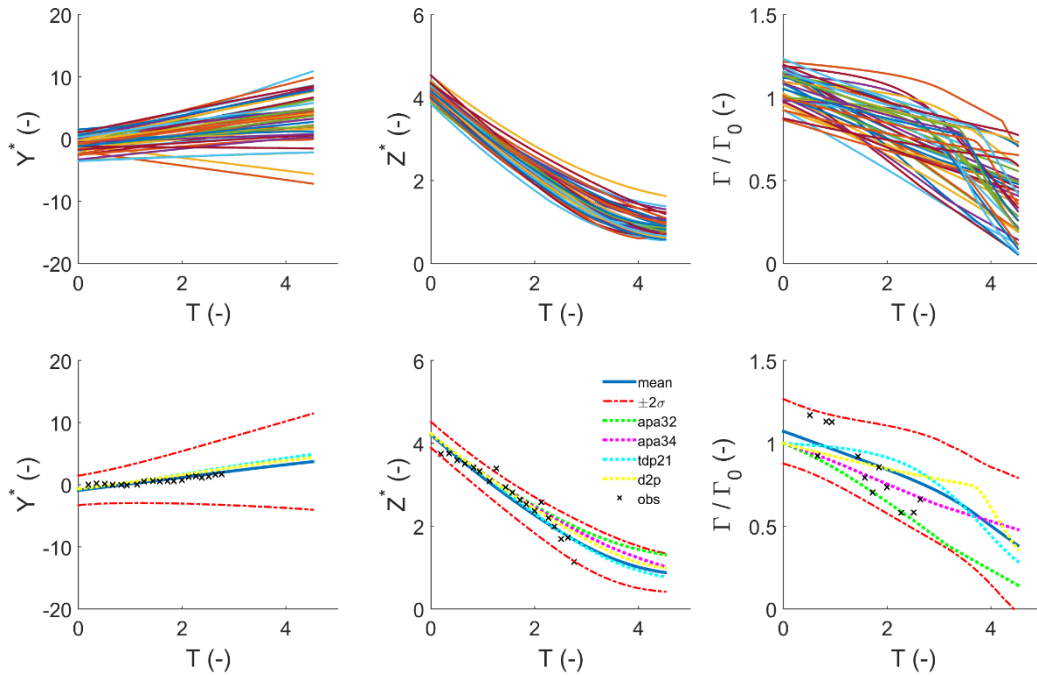

Figure 16.   **Case 1995-08-08-000643.** *The top row shows the results from all simulations. The bottom row shows the mean, and the bounds generated from the Monte-Carlo run along with the results of deterministic simulations of each model.*
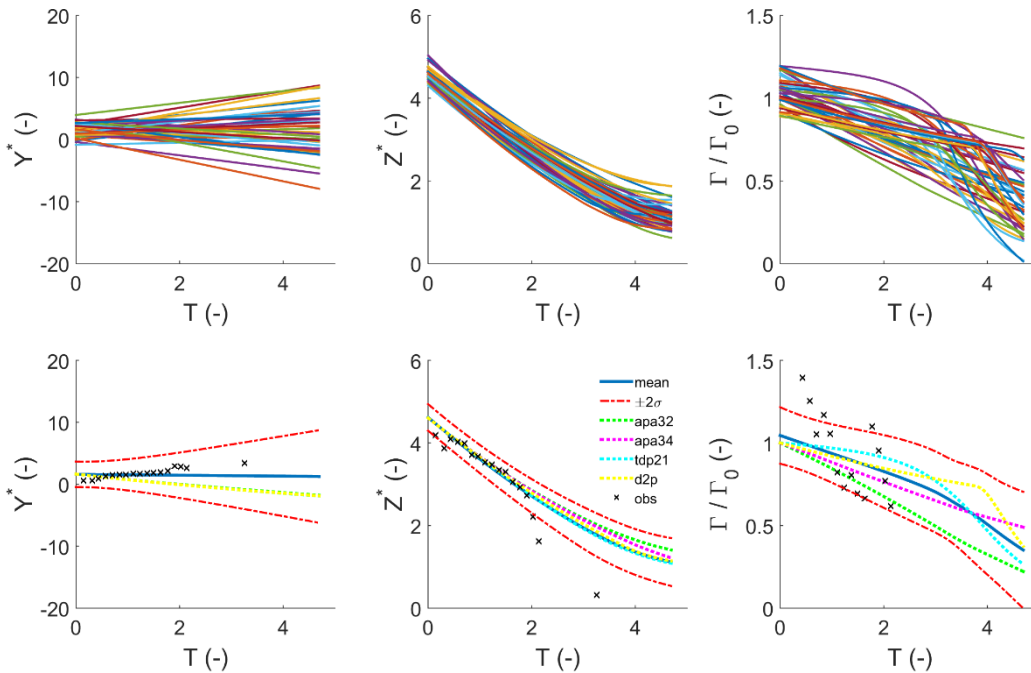
IV.   Comparison of the Methods

In this study, neither the deterministic output of the MCS, nor the probabilistic output of the REA and BMA have been further investigated. Hence a direct comparison, although planned for the future, is only possible for the DEA, the REA and the BMA. Table 12 shows how the approaches perform relatively to the DEA method. Both advanced

American Institute of Aeronautics and Astronautics

approaches outperform the DEA by at least 18 %. In direct comparison of skill, the BMA performs slightly better than the REA, even though the latter delivers better forecasts for y_{luff} and z_{luff}. Comparing the robustness of both advanced methods, we find that the BMA is less sensitive to the training data concerning ensemble parameters. Furthermore the probabilistic envelope of the BMA is very straightforward and considers initial condition uncertainty already in the training phase. In contrast the initial condition uncertainty must be added on top of the envelope of the REA method, which only considers model uncertainty.

Table 12: performance comparison of the individual ensemble methods (median rmse for the test dataset). A positive skill factor indicates, that the respective method outperforms the DEA method on average.

	rmse Γ_{luff}	rmse Γ_{lee}	rmse y_{luff}	rmse y_{lee}	rmse z_{luff}	rmse z_{lee}	skill s
DEA	0.115	0.104	0.829	0.566	0.199	0.167	0.00
REA	0.095	0.097	0.576	0.589	0.137	0.169	0.1872
BMA	0.092	0.093	0.626	0.565	0.139	0.168	0.1866

V.   Summary

This paper investigates the capability to improve wake vortex forecast by combining several independent wake vortex models in a Multi-Model Ensemble. The models that are used in this common study of DLR and NASA were exchanged in an inter-agency cooperation and comprise APA3.2, APA3.4, TDP2.1 (NASA models) and D2P (DLR model). The analysis is based on the following four Multi-Model Ensemble approaches:

- Direct Ensemble Average (DEA)
- Reliability Ensemble Averaging (REA)
- Bayesian Model Averaging (BMA)
- Monte-Carlo Simulation (MCS)

As the independency of the models is crucial for a successful ensemble, the models are compared in an ANOVA analysis. This reveals that APA3.2 and APA3.4 do not differ in a statistically significant way regarding the forecast of the vertical vortex position. Subsequently the different ensemble methods are introduced and then evaluated on the basis of a test dataset (wake vortices of landing aircraft, captured by lidar).

The REA approach (Giorgi and Mearns 2002) utilizes a performance (model bias) and a convergence criterion (distance of model forecast to ensemble mean) to rate the skill of the models. In contrast the BMA method (Hoeting et al. 1999, Raftery et al. 2005) only uses the errors made by the models when they deliver the best forecast among the ensemble for the rating of its members. In addition, the BMA dresses each single model forecast with a pdd that describes the model accuracy. The scoring shows that both the REA and the BMA approach can outperform the model forecast of the best member in the test dataset by at least 1.7 %, and even more if only the z and Γ forecast are regarded. A ranking of the methods on the basis of 200 landings indicates that the advanced approaches outperform the DEA by at least 18% on average (see Table 12). Nevertheless, the DEA approach outperforms the REA method for y_{lee} and the BMA method for z_{lee}. The direct comparison of the advanced approaches reveals that the BMA approach gives slightly better results than the REA approach when looking at the skill factors (see Table 12). The BMA has proven to be more robust. Contrarily to the REA approach it is not necessary to calculate weights that depend on the ambient weather conditions to obtain good results. In addition its uncertainty envelopes are straightforward as they consider initial condition uncertainties by design. They can be chosen according to any desired probability level from the ensemble pdds. At this stage we assume that the standard deviations of the individual predictions do not depend on vortex age. Extending the method by this feature is planned.

Finally a Monte-Carlo Simulation is utilized to study the effect of multiple models on the coverage of possible solutions. The MCS employs model runs based on perturbed initial conditions, such as crosswind speed, turbulence and potential temperature. Each model uses the same set of perturbations, which is randomly generated from pdds that are derived from a set of observations. To determine the success rate of the multi-model MCS the number of observations within the $\pm 2\sigma$ bounds are calculated and compared with the single model MCS. The results suggest that the success rate is improved as the coverage of the solution space can be enhanced by employing multiple models, especially when regarding the Γ-forecast. Runs with different sets of perturbations for each model show similar results.

American Institute of Aeronautics and Astronautics

In contrast to the other methods no deterministic output is determined for the MCS and thus no direct comparison can yet be made concerning the ensemble mean.

In a further study, the amount of observations used for training and scoring shall be increased in order to enhance the significance of the results. In this study only landings with $z_0^* > 1.7$ are regarded for the BMA and REA. Thus future work shall also investigate the performance of ensemble wake vortex predictions for higher initial altitudes. Although both the REA and the BMA offer probabilistic envelopes, this paper focuses on their deterministic performance. Further studies will therefore also concentrate on their probabilistic skill. This way they can be compared with the MCS on the basis of a common dataset.

Appendix A: Evaluation of the Fast-Time Wake Vortex Models

In this section the fast-time models are evaluated using data from four different wake vortex field experiments:

- Memphis 1995 Field Experiment (MEM95). Details on the field experiment are given in Campbell et al. (1997). Wake observations were made using a continuous wave lidar.
- Dallas/Fort Worth 1997 field experiment (DFW97). See Dasey et al. (1997) for description of the field experiment. Wake observations were made using a continuous wave lidar.
- Denver 2003 Field Experiment (DEN03). Details on the field experiment are given in Dougherty et al. (2004). Wake observations were made using a pulsed lidar.
- Oberpfaffenhofen Field Experiment (WakeOP). Details on the field experiment are given in Holzäpfel et al. (2014). Wake observations were made using a pulsed lidar.

The accuracy of model predictions was quantified in terms of *root mean square error* (*Error*$_{rms}$), *mean absolute error* (*Error*$_{mae}$), and *Bias*:

$$Error_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i^{model} - x_i^{obs}\right)^2} \; ; \quad Error_{mae} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i^{model} - x_i^{obs}\right|; \quad Bias = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^{model} - x_i^{obs}\right)$$

The model prediction errors for all MEM95, DFW97, DEN03, and WakeOP cases are given in Tables A.1 through A.4.

Table A.1: Fast-Time Models Evaluation using Memphis 1995 Data – (305 Cases)

Model	Circulation (normalized by Γ_0)			Lateral Transport (normalized by b_0)			Altitude (normalized by b_0)		
	rmse	mae	bias	rmse	mae	bias	rmse	mae	bias
TDP2.1	0.263	0.224	0.029	1.010	0.832	0.095	0.528	0.450	0.072
APA3.4	0.245	0.210	-0.053	0.979	0.807	0.102	0.544	0.463	0.143
APA3.2	0.256	0.221	-0.122	0.996	0.820	0.108	0.545	0.466	0.185
D2P	0.246	0.210	-0.024	1.009	0.833	0.116	0.559	0.474	0.097

American Institute of Aeronautics and Astronautics

Table A.2: Fast-Time Models Evaluation using Dallas 1997 Data – (208 Cases)

Model	Circulation (normalized by Γ_0)			Lateral Transport (normalized by b_0)			Altitude (normalized by b_0)		
	rmse	mae	bias	rmse	mae	bias	rmse	mae	bias
TDP2.1	0.289	0.242	-0.002	0.644	0.519	-0.216	0.290	0.241	0.047
APA3.4	0.273	0.227	-0.076	0.605	0.490	-0.188	0.286	0.238	0.055
APA3.2	0.278	0.233	-0.131	0.608	0.493	-0.188	0.288	0.239	0.036
D2P	0.288	0.241	-0.022	0.607	0.494	-0.189	0.281	0.232	-0.017

Table A.3: Fast-Time Models Evaluation using Denver 2003 Data – (862 Cases)

Model	Circulation (normalized by Γ_0)			Lateral Transport (normalized by b_0)			Altitude (normalized by b_0)		
	rmse	mae	bias	rmse	mae	bias	rmse	mae	bias
TDP2.1	0.296	0.270	0.224	0.729	0.596	-0.142	0.574	0.478	0.070
APA3.4	0.227	0.198	0.152	0.724	0.592	-0.144	0.614	0.512	0.195
APA3.2	0.210	0.177	0.087	0.729	0.593	-0.133	0.622	0.518	0.236
D2P	0.254	0.227	0.184	0.716	0.586	-0.143	0.594	0.494	0.087

Table A.4: Fast-Time Models Evaluation using WakeOP Data – (31 Cases)

Model	Circulation (normalized by Γ_0)			Lateral Transport (normalized by b_0)			Altitude (normalized by b_0)		
	rmse	mae	bias	rmse	mae	bias	rmse	mae	bias
TDP2.1	0.104	0.086	0.035	0.832	0.707	0.447	0.313	0.266	0.157
APA3.4	0.099	0.084	-0.015	0.835	0.710	0.444	0.285	0.243	0.117
APA3.2	0.141	0.120	-0.090	0.865	0.731	0.456	0.288	0.249	0.107
D2P	0.108	0.091	-0.001	0.860	0.729	0.382	0.236	0.203	0.085

American Institute of Aeronautics and Astronautics

Acknowledgments

References

Ahmad, NN, RL VanValkenburg, MJ Pruis, "NASA AVOSS Fast-Time Wake Prediction Models: User's Guide," National Aeronautics and Space Administration, NASA/TM-2014-218152.

Campbell, SD, et al., "Wake Vortex Field Measurement Program at Memphis, TN Data Guide", Lincoln Laboratory, Massachusetts Institute of Technology. Project Report NASA/L-2. 1997.

Corjon, A, and T Poinsot, "Vortex Model to Define Safe Aircraft Separation Distances," *Journal of Aircraft*, Vol. 33, No.3, May-June 1996, pp. 547-553.

Dasey, TJ, et al., "Aircraft Vortex Spacing System (AVOSS). Initial 1997 System Deployment at Dallas/Ft. Worth (DFW) Airport", Lincoln Laboratory, Massachusetts Institute of Technology. Project Report NASA/L-3. 1998.

Dougherty, RP, FY Wang, ER Booth, ME Watts, N Fenichel, RE D'Errico, "Aircraft Wake Vortex Measurements at Denver International Airport," AIAA Paper 2004-2880.

Frech, M and Holzäpfel, F, "Skill of an aircraft wake-vortex model using weather prediction and observation," *Journal of Aircraft*, Vol. 45, 2008, 461–470.

Gerz, T, F Holzäpfel, W Bryant, F Köpp, M Frech, A Tafferner, G Winckelmans, "Research towards a wake vortex advisory for optimal aircraft spacing", *Journal of Applied Meteorology and Climatology*, Vol. 46, 2007, pp. 1913-1932.

Giorgi, F and Mearns, LO, "Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method". *Journal of Climate*, Vol. 15, 2002, pp. 1141–1158.

Greene, GC, "An Approximate Model of Vortex Decay in the Atmosphere", *Journal of Aircraft*, Vol. 23, 1986, pp. 566-573.

Hagedorn, R et al., "The rationale behind the success of multi-model ensembles in seasonal forecasting – i. Basic concept". *Tellus*, Vol. 57 A, 2005, pp.219–233.

Hinton, DA, "Aircraft Vortex Spacing System (AVOSS) Conceptual Design", NASA Technical Memorandum NASA-TM-110184, 1995.

Han, J, SP Arya, S Shen, Y Lin, "An Estimation of Turbulent Kinetic Energy and Energy Dissipation Rate Based on Atmospheric Boundary Layer Similarity Theory", NASA Contractor Report NASA-CR-2000-210298. 2000.

Han J, Y Lin, SP Arya, FH Proctor, "Numerical Study of Wake Vortex Decay and Descent in Homogeneous Atmospheric Turbulence," *AIAA Journal,* Vol. 38, 2000, pp. 643-656.

Holzäpfel, F, "Probabilistic Two-Phase Wake-Vortex Decay and Transport Model," *Journal of Aircraft*, Vol. 40, 2003, pp. 323-331.

Holzäpfel, F et al., "Impact of wind and obstacles on wake vortex evolution in ground proximity." AIAA Paper 2014-2470.

Holzäpfel, F and Steen, M, "Aircraft wake-vortex evolution in ground proximity: Analysis and parameterization." *AIAA Journal*, Vol. 45, 2007, pp. 218–227.

Holzäpfel, F., "Effects of Environmental and Aircraft Parameters on Wake Vortex Behavior", *Journal of Aircraft*, Vol. 51, 2014, pp. 1490-1500, doi: 10.2514/1.C032366..

Hoeting et al., "Bayesian Model Averaging: A Tutorial", Statistical Science, Vol. 14, No. 4, 1999, pp. 382-417.

Joseph, R, T Dasey, R Heinrichs, "Vortex and Meteorological Measurements at Dallas/Ft. Worth Airport," AIAA Paper 1999-0760.

American Institute of Aeronautics and Astronautics

Köpp, F et al., "Comparison of wake-vortex parameters measured by pulsed and continuous-wave lidars." *Journal of Aircraft*, Vol. 42, 2005, pp.916–923.

Perry, RB, DA Hinton, and RA Stuever, "NASA Wake Vortex Research for Aircraft Spacing," AIAA Paper 1997-0057.

Proctor, FH, "The Terminal Area Simulation System / Volume 1: Theoretical Formulation", NASA Technical Report 4046. 1987.

Proctor, FH, "The NASA-Langley Wake Vortex Modelling Effort in Support of an Operational Aircraft Spacing System", AIAA-98-0589.

Proctor, FH, DW Hamilton, and J Han, "Wake Vortex Transport and Decay in Ground Effect: Vortex Linking with the Ground," AIAA-2000-0757.

Proctor, FH, DW Hamilton, GF Switzer, "TASS Driven Algorithms for Wake Prediction," American Institute of Aeronautics and Astronautics, AIAA-2006-1073.

Proctor, FH, "Interaction of Aircraft Wakes from Laterally Spaced Aircraft," American Institute of Aeronautics and Astronautics, AIAA-2009-343.

Proctor, FH, "Evaluation of Fast-Time Wake Vortex Prediction Models," AIAA Paper 2009-0344.

Pruis, MJ, DP Delisi, NN Ahmad, "Comparisons of Crosswind Velocity Profile Estimates Used in Fast-Time Wake Vortex Prediction Models," AIAA Paper 2011-1002.

Pruis, MJ and DP Delisi, "Comparison of Ensemble Predictions of a New Probabilistic Fast-Time Wake Vortex Model and Lidar Observed Vortex Circulation Intensities and Trajectories," AIAA Paper 2011-3036.

Pruis, MJ, and DP Delisi "Correlation of the Temporal Variability in the Crosswind and the Observation Lifetime of Vortices Measured with a Pulsed Lidar," AIAA Paper 2011-3199.

Pruis, MJ, and Delisi, DP, "Assessment of fast-time wake vortex prediction models using pulsed and continuous wave lidar observations at several different airports." AIAA Paper, 2011-3035.

Raftery, A, Gneiting, T, Balabdaoui, F, and Polakowski, M, "Using Bayesian Model Averaging to calibrate forecast ensembles." *Monthly Weather Review*, Vol. 133, 2005, pp.1155–1174.

Robins, RE, and DP Delisi, "NWRA AVOSS Wake Vortex Prediction Algorithm Version 3.1.1," NASA CR 2002-211746.

Sarpkaya, T, "New Model for Vortex Decay in the Atmosphere," *Journal of Aircraft*, Vol. 37, 2000, pp. 53-61.

Sarpkaya, T, RE Robins, and DP Delisi, "Wake-Vortex Eddy-Dissipation Model Predictions Compared with Observations," *Journal of Aircraft*, Vol. 38, 2001, pp. 687- 692.

Weigel, AP et al., "Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?" *Quarterly Journal of the Royal Meteorological Society*, Vol. 134, 2008, pp.241–260.

American Institute of Aeronautics and Astronautics