Multi-Model Ensembles for Upper Atmosphere Models

Elvidge, S.; Granados, S. R.; Angling, M. J.; Themens, D. R.; Wood, A. G.; Brown, Matthew

DOI:
10.1029/2022SW003356

AGU ADVANCING EARTH AND SPACE SCIENCE

Space Weather®

RESEARCH ARTICLE
10.1029/2022SW003356

Multi-Model Ensembles for Upper Atmosphere Models

S. Elvidge[1], S. R. Granados[1], M. J. Angling[2], M. K. Brown[1], D. R. Themens[1], and A. G. Wood[1]

[1]Space Environment and Radio Engineering Group (SERENE), University of Birmingham, Birmingham, UK, [2]In-Space Missions Ltd, Alton, UK

Key Points:

- Multi-model ensembles (MMEs) are used to reduce the error in specifying the thermosphere
- The MME performs better than any individual model in all test scenarios
- A non-negative least squares weighting for the MME reduces the error by 68% at solar minimum and 50% at solar maximum

Correspondence to:

S. Elvidge,
s.elvidge@bham.ac.uk

Citation:

Elvidge, S., Granados, S. R., Angling, M. J., Brown, M. K., Themens, D. R., & Wood, A. G. (2023). Multi-model ensembles for upper atmosphere models. *Space Weather*, *21*, e2022SW003356. https://doi.org/10.1029/2022SW003356

Received 22 NOV 2022
Accepted 28 FEB 2023

Abstract Multi-model ensembles (MMEs) are used to improve the forecasts of thermospheric neutral densities. A variety of algorithms for constructing the model weights for the MMEs are described and have been implemented including: performance weighting, independence weighting, and non-negative least squares. Using both empirical and physics-based models, compared against in situ Challenging Minisatellite Payload (CHAMP) observations, the skill of each MME weighting approach has been tested in both solar minimum and maximum conditions. In both cases the MME performs better than any individual model. A non-negative least squares weighting for the MME on a set of bias corrected models provides a 68% and 50% reduction in the mean square error compared to the best model (Jacchia-Bowman 2008) in the solar minimum and maximum cases, respectively.

Plain Language Summary Combining multiple models of the neutral upper atmosphere (thermosphere) can lead to the cancellation of errors and improved short-term forecasts of the environment. In this paper a number of different methods for creating these "multi-model ensembles" (MMEs) are investigated, varying how the different models in the comparison are weighted and combined. Using both statistical and first-principles models and compared to observations from the Challenging Minisatellite Payload (CHAMP) satellite, the skill of each MME approach has been tested in both solar minimum and maximum conditions. In both cases the MME performs better than any individual model. The best performing combination makes a 68% reduction in the mean square error compared to the best individual model at solar minimum and a 50% improvement at solar maximum.

1. Introduction

1.1. Background

Accurately propagating satellite orbits requires knowledge of the forces acting on the satellite. For satellites in low Earth orbit (less than 1,000 km), forces include terrestrial gravity, solar radiation pressure, lunar and solar gravity and drag caused by the atmosphere (Eshagh & Najafi Alamdari, 2007). The drag force increases dramatically as a satellite's altitude decreases and becomes significant below approximately 600 km (Fortescue et al., 2011). However, there are large uncertainties in modeling the magnitude of the drag acting on a satellite. To do so requires an understanding of the thermospheric mass density, winds and the satellite's ballistic coefficient. The largest contribution to error in the forecasting of satellite positions is specification of thermospheric density (Mehta et al., 2018), although for tumbling or complex geometries, the errors in the ballistic coefficient can be a substantial contribution.

Currently a variety of mathematical models are used to provide estimates of the density. Empirical models are often used by satellite operators. They are fitted to measurements of thermospheric parameters; however such measurements are sparse. In particular, there are very few measurements between 100 and 250 km because balloons cannot reach these heights and satellites re-enter too quickly for any long term study. Fabry-Perot Interferometers can be used to measure wind between 220 and 600 km (Titheridge, 1995) and meteor radars can measure wind, as well as temperature and pressure, between 80 and 100 km (John et al., 2011; Reid et al., 2018).

Physics-based models solve the equations which describe the physical processes in the thermosphere. Initially the atmospheric density, wind and temperatures are generally provided by empirical models, but a "spin-up" time is used for the results to stabilize. The spin-up time can be reduced in subsequent model runs by using previous output from the model. Neutral and ion species production is then calculated via chemical reaction equations and using solar X-rays and EUV conditions. Ion transportation and recombination are also considered. The initial and boundary conditions, as well as proxies for solar activity, are the main drivers for the models. There are a number of approaches to modeling the physics of the thermosphere, which rely on different numerical methods

(Augenbaum, 1984; Bott, 1989; Purnell, 1976), and thus exhibit different levels of complexity and use a variety of inputs. Model developer choices about the solver and the selection of boundary conditions leads to differences in the outputs from models.

1.2. Multi-Model Ensembles

A multi-model ensemble (MME) is a combination (usually weighted) of individual models (Murray, 2018; Thompson, 1977). Ideally the models should have independent errors and the improved performance of the MME arises from the errors partially canceling (Hagedorn et al., 2005). Tracton and Kalnay (1993) showed the utility of MMEs in one of the first operational MME weather forecasts, and also demonstrated skill in longer-term forecasts. Elvidge (2014) and Elvidge et al. (2016) demonstrated for the first time the skill from using MMEs for upper atmosphere forecasts. The result has been further demonstrated by combining the four Utah State University Global Assimilation of Ionospheric Measurements (USU-GAIM) models (Schunk et al., 2016).

A key question when using an MME is how the models should be combined. Elvidge et al. (2016) used both an equal weighting scheme and a scheme where the weights were the inverse of the mean square error (MSE) of the models used to create the MMEs. That work showed that 6 hr forecasted densities of the thermosphere had a 60% reduction in the root mean square error when using an MME. To further investigate MMEs in the thermosphere Elvidge et al. (2016) recommended that:

- A "training" data set should be used for the weighting scheme rather than the testing data set
- More variety of weighting methods should be included
- Longer test scenarios should be used to reduce the uncertainties in the statistics.

This work addresses those recommendations.

2. Method

2.1. Observations

This work uses data from the Challenging Minisatellite Payload (CHAMP) satellite (Reigber et al., 2002). CHAMP was operational from July 2000 to September 2010. During this time the CHAMP orbit degraded from an altitude of 454 to 296 km due to atmospheric drag. One of its primary missions was to precisely measure the terrestrial gravity field which required a very accurate accelerometer. Thermospheric total mass densities have been estimated from the CHAMP accelerometer data (Sutton, 2009). This is the data set as was used in Elvidge et al. (2016); however, since that study, the CHAMP drag coefficient and surface area has been re-analyzed using higher fidelity satellite geometry models and more advanced drag coefficient estimation (Mehta et al., 2017). This has resulted in a 20% reduction of the estimated densities. This work uses the re-analyzed data, but it should be noted that the empirical models JB2008 and DTM-2013 are fitted with the older data (see Section 3).

The CHAMP data has a high sampling rate along its orbit (10 s). However, the local solar time varies by only a few seconds every orbit. Therefore, over the course of a month, CHAMP would only sample approximately 2.8 hr of local solar time (Häusler et al., 2010). Even so, structures in the neutral density such as traveling atmospheric disturbances and the midnight density maximum can be seen (Emmert, 2015). However the limit of the data coverage does impact the overall assessment of this method. The accelerometer-derived neutral density observations have an estimated mean error of 10.6% (Sutton, 2008).

2.2. MME Weighting Methods

In this paper the performance of six different weighting schemes that can be used to combine models are tested. These are equal weights (EW), performance weights (PW), PW with bias removed (PWB), Reliability Ensemble Averaging (REA), Independence Weighting (IW), and Non-Negative Least Squares Regression (NNLS). Each

scheme is described in the following sections. In each case the MME is formed as a weighted combination of the input models:

$$M(x,t) = \sum_{k=1}^{N} R_k(x,t) Z_k(x,t), \tag{1}$$

where $M(x, t)$ is the MME for each point x at time t, $R_k(x, t)$ is the weight for each model k which could also vary in time and space, $Z_k(x, t)$ is each model output, and N is the number of models.

2.2.1. Equal Weighting

Constructing appropriate model weights can be difficult given small sample sizes and available data (Kharin & Zwiers, 2002). As such it has been argued that the only way to generate a good MME for small data sets is by taking the ensemble mean (Hagedorn et al., 2005). Though simple, this method has been shown to produce good results in the thermosphere (Elvidge et al., 2016) and more broadly in climatology (Barnston et al., 2003; Palmer et al., 2004; Weisheimer et al., 2009). The MME weights are given by:

$$R_k = \frac{1}{N}. \tag{2}$$

2.2.2. Performance Weights

A performance weighting scheme uses a measure of model skill to weight the models so that the best performing model (against a representative data set) has the highest weight. The performance weighting used in this work is a modified version of that described by Rozante et al. (2014). It uses the MSE as the skill measure to weight the models. The weights then have the value:

$$R_k = \frac{\sum_{k=1}^{N} MSE_k}{MSE_k}, \tag{3}$$

where

$$MSE_k = \frac{\sum_{i=1}^{Np} \left(Y_k^i - X^i\right)^2}{Np}, \tag{4}$$

where Y_k^i is a model prediction in the training period, X^i is a data point in the training period and Np is the number of points.

2.2.3. Performance Weights With Bias Removed

Some models show a good amount of skill in terms of the correlation, showing that they model the thermospheric response well. However, the MSE may still be large, and a reason for this can be model biases. The bias is the difference of the average density of each model in the training period and the average density of the data set:

$$B_k = \frac{\sum_{i=1}^{Np} Y_k^i - X^i}{Np}, \tag{5}$$

so that

$$Y_k'^i = Y_k^i - B_k \tag{6}$$

and

$$R_k = \frac{\sum_{k=1}^{N} MSE'_k}{MSE'_k}, \tag{7}$$

where

$$MSE'_k = \frac{\sum_{i=1}^{Np} \left(Y_k'^i - X^i\right)^2}{Np}, \tag{8}$$

If the bias from the training data is removed before MSEs are taken, a potentially better representation of model skill can be achieved. The MSE of an unbiased model is equal to the variance, so this is effectively a variance weighting. The bias is then pre-removed from the validation data set before averaging the models. This assumes the bias does not change between testing and validation, which for a short time should be a reasonable approximation.

2.2.4. Reliability Ensemble Averaging

Elvidge et al. (2016) suggested the use of REA to estimate the ensemble weights. REA is used in terrestrial weather climatology to infer the unknown future performance of the model from its previous performance and in comparison to the other model's predictions (Giorgi & Mearns, 2002). The weighting process involves calculating the following quantity:

$$R_k^j = \min\left(1, \left\{\left[\frac{\epsilon}{abs(B_k)}\right]^m \left[\frac{\epsilon}{abs\left(D_i^j\right)}\right]^n\right\}^{\left[\frac{1}{mn}\right]}\right). \tag{9}$$

The R_k^j are weights per model k and validation point j, the ϵ are estimations of the data set's variability which could be the range or the standard deviation of the data (a constant value of 1×10^{-12} was used in this work as an estimation), B_k is the bias of the model calculated against previous data, D_k^j are distances from the models to the weighted multi-model average, $\widetilde{Y^j}$, given by

$$\widetilde{Y^j} = \frac{\Sigma_k R_k^j Y_k^j}{\Sigma_k R_k^j}, \tag{10}$$

and m and n allow for separate weightings of the bias and the distances (usually $m = n = 1$ (Giorgi & Mearns, 2002)). This is a circular definition since R_k is defined in terms of the distance from $\widetilde{Y^j}$ in Equation 9 so an iterative procedure is used to find the weights and is usually complete within a few cycles. The weights are calculated using Equation 9 then a new average is calculated using Equation 10 until a weight reaches a value of one (Giorgi & Mearns, 2002). This could be useful in storm time when little is known about the storm. It relies on the model average which, a better estimate of a model's reliability than its prestorm bias.

2.2.5. Independence Weighting

Model independence is a critical requirement for an MME to work (Elvidge, 2014). It may be the case that a set of models are not independent and share a lot of their structure with each other. The "independence weighting" approach aims to take this into account. To determine the level of independence between models first each has its bias removed. Ideally this de-biased time series should have Gaussian errors and the covariance between different independent model errors would be zero. In practice often these errors do have some covariance. A covariance matrix of these errors, for each of the different bias corrected models is constructed, and weights are produced based on the variance between model and data, and the amount of covariance between the models (Bishop & Abramowitz, 2013):

$$R_k = \frac{\mathbf{A}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{A}^{-1}\mathbf{1}}, \tag{11}$$

where $\mathbf{1}$ is a vector of all 1's and $\mathbf{A}$ is the model difference covariance matrix. This system can produce negative weights which is meaningless. So the method is adjusted to give only positive weights (Bishop & Abramowitz, 2013):

$$\tilde{R}_k = \frac{R_k^T - \mathbf{1}^T \min(R_k)}{1 - k\min(R_k)}. \quad (12)$$

A consequence of this is that one model always has zero weight, and is therefore excluded from the weighting. This method allows the use of different versions of the same model since independence is no longer a concern, potentially allowing similar models of the different versions/generations and formulations to be used.

2.2.6. Non-Negative Least Squares

Non-Negative Least Squares is a simple constrained regression which does not allow the coefficients to become negative. Specifically it finds the coefficients R_k such that

$$\arg\min_{R_k} \| Z_k R_k - M \|_2^2 \quad \text{subject to } R_k \geq 0 \quad (13)$$

where $\| \cdot \|_2$ is the Euclidean norm (Bro & De Jong, 1997). The regression is performed on the training data set.

3. Models

In this paper six models have been used to create the MME. Three of the models: NRLMSISE-00 (Picone et al., 2002), the Thermosphere Ionosphere Electrodynamic General Circulation Model (TIE-GCM) (Qian et al., 2014) (version 1.95 was used in Elvidge et al. (2016) whilst version 2.0 is used here) and the Global Ionosphere-Thermosphere Model (GITM) (Ridley et al., 2006) (updated since then) were used in Elvidge et al. (2016) (refer to that paper for a brief description of the models, or to the references for a detailed description). GITM and TIE-GCM were both run at 5° × 5° resolution. Additionally the Coupled Thermosphere-Ionosphere Plasmasphere Electrodynamics (CTIPe) (Codrescu et al., 2012; Millward et al., 1996), Jacchia-Bowman 2008 (JB2008) (Bowman et al., 2008) and the Drag Temperature Model 2013 (DTM-2013) (S. Bruinsma, 2015) are used in this paper. A summary of the differences between the empirical models used in this work are shown in Table 1 of Emmert (2015) whilst Table 2 of the same paper highlights the difference between the physics-based models.

3.1. CTIPe

The Coupled Thermosphere-Ionosphere-Plasmasphere-electrodynamics model (CTIPe) has been developed at the National Oceanic and Atmospheric Administration (NOAA). It is a physics-based model with a fixed resolution of 18 cells in longitude, 90 in latitude and 15 vertical pressure levels. These values are due to the smaller scales of spatial variation in latitude compared to longitude. The model assumes hydrostatic equilibrium (as TIE-GCM from Elvidge et al. (2016) does). As well as F10.7, CTIP uses hemispheric power in 12 min intervals. The model was run on request at the Community Coordinated Modeling Center (CCMC) website and automatically interpolated to CHAMP paths on the website (Codrescu et al., 2012).

3.2. JB2008

Jacchia-Bowman 2008 (JB2008) has been developed by Space Environment Technologies (SET) and is an empirical thermospheric density model (Bowman et al., 2008). It is based on the previous JB2006 and the original Jacchia diffusion equations (Bowman et al., 2008; Jacchia, 1977). The model uses four solar proxies (computed from in-orbit sensors) as well as disturbance storm time index (Dst) data (a measure of geomagnetic activity) (Tobiska et al., 2009). The model has been validated using derived density data from satellite drag on a range of satellites been 175 and 1,000 km.

3.3. DTM-2013

The Drag Thermosphere Model 2013 (DTM-2013) is a semi-empirical model which describes thermospheric temperature, density, and composition. The model has been developed by the Centre National d'Etudes Spatiales (CNES) and has a long development history starting with DTM-78 (Barlier et al., 1978). DTM derives its densities and temperatures from satellite drag data and was the first model to include the high-accuracy accelerom-

Table 1
Labels in the Modified Taylor Diagrams

MME	Abbreviation
Equal	EW
Performance weighting	PW
Performance weighting with bias subtracted	PWB
Reliability ensemble averaging	REA
Non-negative least squares	NNLS
Independence weighting	IW

eter data from the CHAMP and GRACE satellite missions (S. L. Bruinsma et al., 2004). Recent developments of the model include GOCE satellite data from 270 km to improve specification of the lower thermosphere and use of F30 (30 cm radio flux) instead of the F10.7. These updates have shown to increase the performance of the model with regards to specifying thermospheric density (S. Bruinsma, 2015). It uses *am* instead of *ap* for modeling geomagnetic storm modeling.

4. Test Scenarios

4.1. Solar Minimum Scenario

The test scenario used in this work is an extension of the first test scenario in Elvidge et al. (2016), a 20 days long run from 18 August 2009 (Elvidge et al. (2016)'s test scenario started on 28 August 2009). The MME weighting schemes which require training, are trained on the 20 day period before this (from July 29). The test scenario time period contains one geomagnetic storm, Figure 1, and during the rest of the month the geomagnetic conditions are quiet.

The F10.7 only varies between 67 and 76 flux units. The extremely low solar minimum of 2008–2009 presents a significant modeling challenge since the F10.7 values have been shown to not represent the correct thermospheric conditions (Bilitza et al., 2017; Solomon et al., 2010). However at solar minimum, internal, and external dynamics, rather than solar drivers, dominate the evolution of the thermospheric densities. It is expected that the greatest differences between the tested models will be evident at these times (Elvidge et al., 2016). A sample timeseries of a physics-based and empirical model and the data for this period is shown in Figure 2. Recall that the average mean error of the observations is approximately 10.6% (Sutton, 2008), which is shown as error bars around the CHAMP data points. Whilst the errors are not insignificant they are smaller than the differences between the models. The fast periodicity of the data is due to the CHAMP satellite completing one orbit every 90 min and each point being 15 min apart.

4.2. Solar Maximum Scenario

The second test scenario is a typical 30 days solar maximum period from 2002, using a 30 days training window. The F10.7 varies between 135 and 240, with some significant spikes in *ap* (Figure 3).

5. Results

5.1. Introduction

The various upper atmosphere models, and the different MME approaches (whose weights are calculated on the training periods) have been run for the test scenarios. These forecasts are compared to the derived CHAMP neutral densities which have an estimated mean error of 10.6% (Sutton, 2008), and whilst these results cover month long scenarios that only represents approximately 2.8 hr of local solar time coverage (Häusler et al., 2010).

Table 2
Weighting of the Different Models in 2009

	NRLMSISE-00	JB2008	DTM2013	TIE-GCM	GITM	CTIPe
Equal weighting	0.17	0.17	0.17	0.17	0.17	0.17
Performance weighting	0.12	0.47	0.19	0.08	0.10	0.04
Performance weighting with bias subtracted	0.24	0.25	0.21	0.07	0.11	0.12
Reliability ensemble averaging[a]	–	–	–	–	–	–
Non-negative least squares	0.10	0.10	0.35	0.11	0.21	0.13
Independence weighting	0.00	0.00	0.59	0.04	0.28	0.09

[a]Weights vary over time.

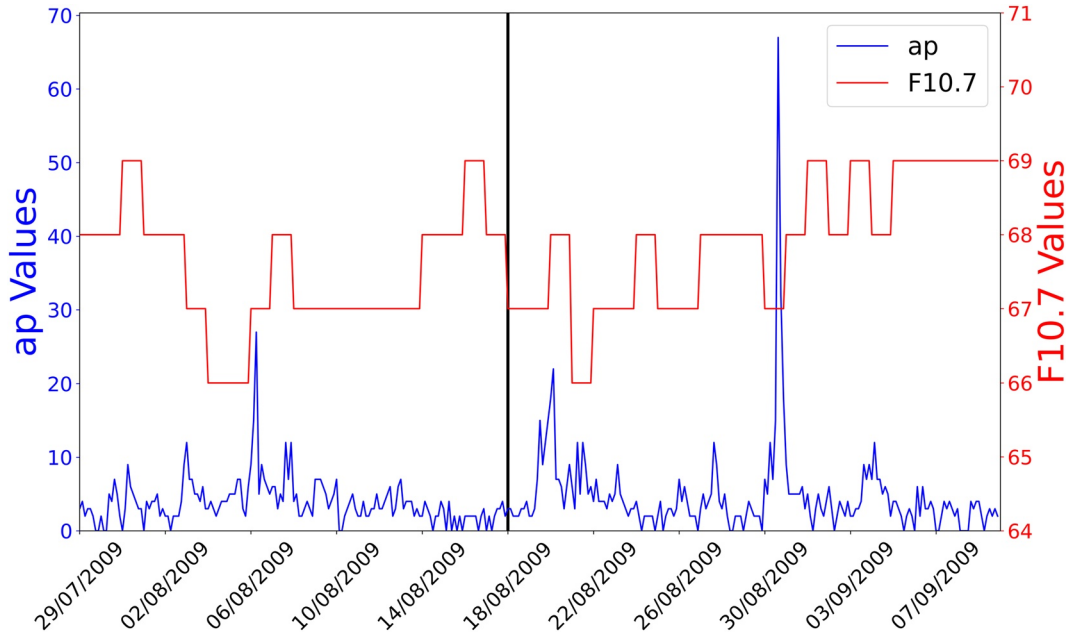

Figure 1. *ap* (blue) and F10.7 (red) for the test scenario and training period which runs from 29 July to 8 September 2009. Training period is before 18 August (black line), values after this are used for validation. The large spike in *ap* is associated to a geomagnetic storm.

The models are compared using modified Taylor diagrams (Elvidge et al., 2014). To read such a diagram (e.g., Figure 4): the radial distance of a data point from the origin is the models normalized standard deviation (here they are normalized by the standard deviation of the observation), and the azimuthal angle corresponds to the correlation between the model and observation. The dashed line (marked with a star) shows the normalized standard deviation of the observation (i.e., unity). The dotted lined semicircles, originating from the intersection of the observed standard deviation (dashed line) and the horizontal axis, show contours of the standard deviation of the model error. Finally, the (normalized) MSE between the model and observation time series can be found by adding, in quadrature, the standard deviation of the model error and model bias (model minus observations) which is shown by the color bar. The normalization factors have been included in the top right of the diagram and can be used to revert any factor to its original value.

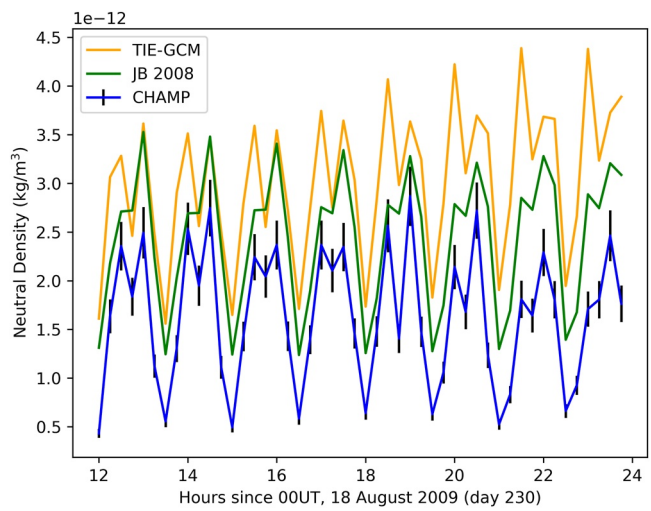

Figure 2. Sample timeseries of neutral density from the solar minimum scenario for CHAMP (blue, with error bars shown in black), TIE-GCM (orange) and JB2008 (green).

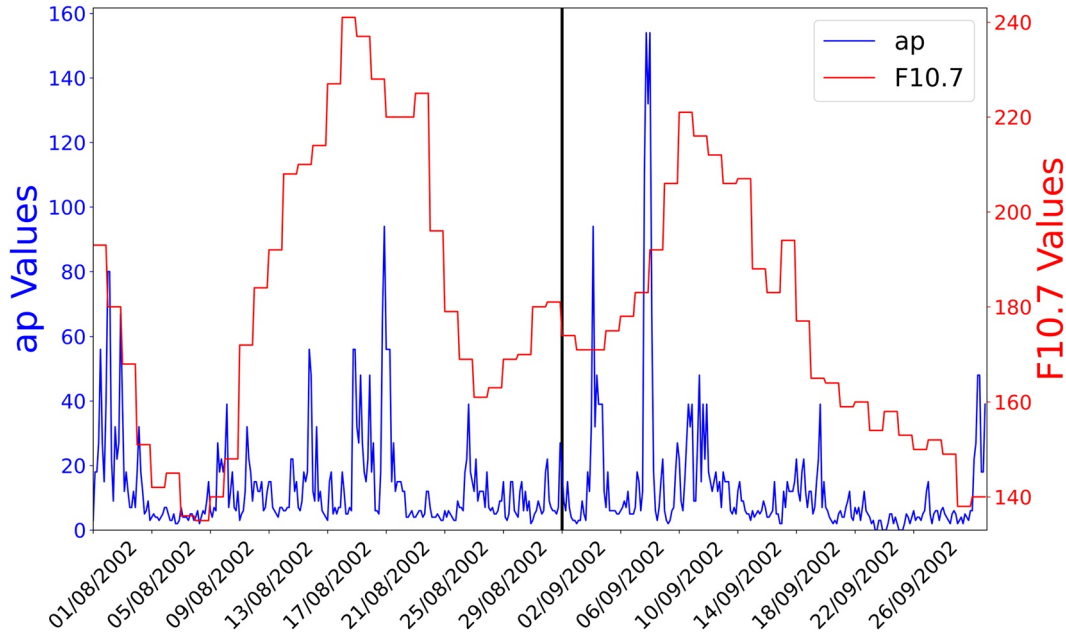

Figure 3. *ap* (blue) and F10.7 (red) for the test scenario and training period which runs from 1 August to 30 September 2002. The training period is before 31 August (black line), values after this are used for validation. The large spikes in *ap* are associated to geomagnetic storms.

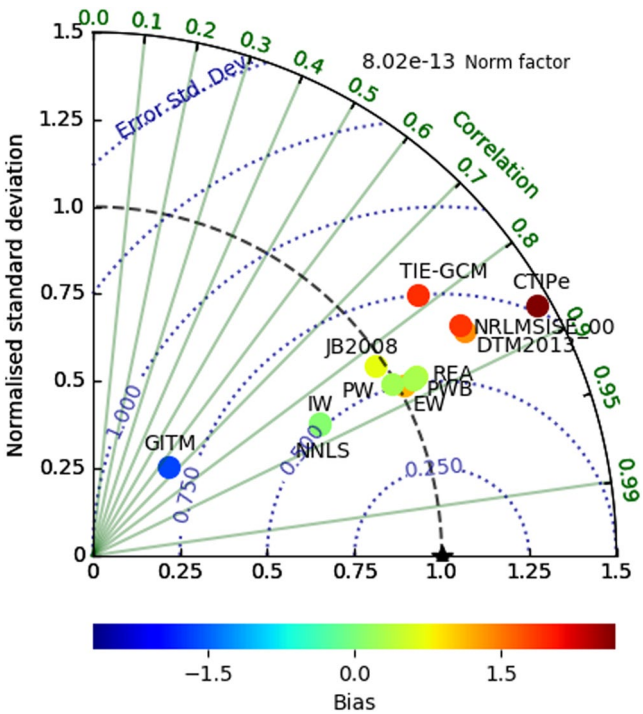

Figure 4. Modified Taylor Diagram for 20 days from 18 August 2009 with 20 days training period. Data from the CHAMP satellite. The diagram shows the correlation to the data, normalized standard deviation, and bias. The label expansions are shown in Table 1.

5.2. Solar Minimum Scenario

Figure 4 shows a modified Taylor diagram of the solar minimum test scenario. It can be seen that TIE-GCM and CTIPe have large positive biases and normalized standard deviations greater than 1 compared to the CHAMP observations. Whilst the other physics-based model, GITM, underestimates the range of observations (normalized standard deviation significantly less than 1) and is negatively biased.

TIE-GCM and GITM have a lower correlation to the data than the empirical models, whilst CTIPe is similar. NRLMSISE-00 has a large positive bias and variance although with a moderately better correlation than TIE-GCM. DTM performs similarly to NRLMSISE, but with a slightly smaller bias. Overall JB2008 performs the best of any individual model. Of the MME approaches, the simple equally weighted ensemble leads to a greater correlation with the data compared to any individual model and accurate variance. It is a common theme for thermospheric models to overestimate the neutral density, but GITM here has a low bias which improves the bias of the equally weighted ensemble. REA and PWB have near-zero biases due to the bias correction, this shows that the biases in these models varies over timescales longer than a month. The non-negative least squares and independence ensembles perform similarly to the others but with a lower variance. The model with the highest MSE is TIE-GCM at 2.80×10^{-24}, and the lowest is JB2008 with an MSE of 5.35×10^{-25}, significantly lower than all the others in this regime. Of the MMEs the highest MSE is equal weighting with an MSE of 1.03×10^{-24} and the lowest is the non-negative least squares with an MSE of 1.73×10^{-25}. The maximum drop in MSE therefore is 94%, and a 68% from the best model.

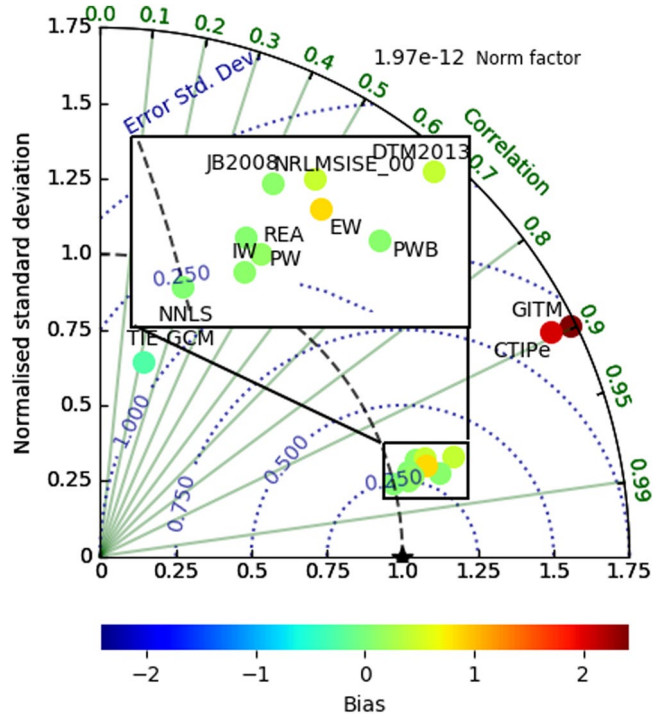

Figure 5. Modified Taylor Diagram for 30 days validation from 31 August with 30 days training period. Data from the CHAMP satellite. The diagram shows the correlation to the data, normalized standard deviation, and bias. The timeseries have not been binned.

The weightings of each model, for the different schemes used here, are shown in Table 2. After removing the bias the weights allocated to NRLMSISE-00 become higher and the weights for JB2008 become lower. The regression and independence weightings both favor the physical models more heavily at the expense of JB2008.

5.3. Solar Maximum Scenario

Figure 5 shows a modified Taylor diagram for the individual models and MME results for the solar maximum test scenario. The empirical model performance is superior. GITM again has a high bias and variance along with CTIPe.

The MSE of JB2008 is 4.53×10^{-25} and the worst model is GITM with an MSE of 2.62×10^{-23}. The physics models generally performed worse here than in the 2009 solar minimum test. GITM has a positive bias and a greater than 1 normalized standard deviation in contrast to the solar minimum test scenario when it is less than 1 and negatively biased. In this case TIE-GCM has a negative bias, again in contrast to 2009. TIE-GCM also has a small correlation, implying it had trouble producing the correct features in the neutral density field, whereas GITM produced a very high correlation. The best MME was again a non-negative least squares with an MSE of 2.35×10^{-25}, while the equally weighted ensemble had an MSE of 3.28×10^{-24}. The highest improvement is 99%, and the improvement from JB2008 is 49%. The weightings are shown in Table 3. It can be seen that the physics-based models are weighted less heavily than the empirical models. NRLMSISE-00 is weighted more heavily than in the non-negative least squares. The equally weighted MME did not have a lower MSE than the best performing model in either circumstance.

Table 3
Weighting of the Different Models in 2002

	NRLMSISE-00	JB2008	DTM2013	TIE-GCM	GITM	CTIPe
Equal weighting	0.17	0.17	0.17	0.17	0.17	0.17
Performance weighting	0.11	0.52	0.30	0.06	0.01	0.01
Performance weighting with bias subtracted	0.28	0.29	0.27	0.03	0.06	0.07
Reliability ensemble averaging[a]	–	–	–	–	–	–
Non-negative least squares	0.35	0.44	0.16	0.00	0.04	0.00
Independence weighting	0.29	0.46	0.15	0.00	0.07	0.04

[a]Weights vary over time.

6. Conclusions

MMEs have been shown to improve the MSE of upper atmosphere forecasts. They rely on a spread of values around the true value to approximate it. Upper atmosphere models tend to be biased and for satellite predictions this is the most important statistical parameter, since the bias leads to a consistent deviation away from the true satellite track. Models (and MMEs) therefore need some kind of bias correction. Efforts like HASDM (Storz et al., 2005) where the biases of a thermospheric model are corrected by data assimilation can reduce them to near zero, and lead to vastly improved satellite prediction capabilities. However MMEs offer an opportunity to de-bias the model output simply, without the need for a computationally expensive data assimilation system, and can be used during forecasts where data is unavailable. A number of different MME methodologies have been described and compared here which can broadly be used throughout space weather (not just in the context of thermospheric density specification). If deploying such a system in an operational setting we would recommend that weights are calculated on a "rolling" 1-month basis (if not using REA which, by definition, varies over time).

This paper has investigated the recommendations of Elvidge et al. (2016) to improve our understanding of the use of MMEs in the thermosphere. Training datasets have been used to calculate the individual model weights, and a greater variety of weighting schemes have been used. The testing scenarios have also been extended to reduce the uncertainties in the statistics. In both the solar maximum and minimum test scenarios the MME performs better than any individual model compared in this study, within the confines of only using CHAMP data. Whilst many of the MME weighting methods perform similarly, overall a non-negative least squares weighting on bias corrected models gives the largest reduction in error. In the solar minimum case this is a 68% reduction in the MSE from the best individual model (Jacchia-Bowman 2008 [JB2008]) and a 50% reduction in the solar maximum case, again compared to JB2008.

Data Availability Statement

The CHAMP data were collected from http://tinyurl.com/densitysets as provided by Mehta et al. (2017). TIE-GCM is developed by NCAR and is available at http://www.hao.ucar.edu/modeling/tgcm/tie.php. NRLMSISE-00 was developed by NRL and is available via the Community Coordinated Modeling Center (CCMC) at https://kauai.ccmc.gsfc.nasa.gov/instantrun/msis/. GITM was developed by Aaron Ridley at the University of Michigan and is available at https://github.com/aaronjridley/GITM. CTIPe was developed at NOAA and was run via the "runs-on-request" system on CCMC https://ccmc.gsfc.nasa.gov/models/modelinfo.php?model=CTIPe. JB2008 is provided by Space Environment Technologies from https://sol.spacenvironment.net/jb2008/code.html and finally DTM-2013 was provided by Dr. Sean Bruinsma, CNES, Space Geodesy Office. DTM-2020 is available from https://github.com/swami-h2020-eu/mcm.

Acknowledgments
SE and MKB are supported by the UK Space Weather Instrumentation, Measurement, Modelling and Risk (SWIMMR) Programme, National Environmental Research Council (NERC) Grant NE/V002708/1. DRT's contribution to this work is supported in part through Canadian Space Agency Grant 21SUSTCHAI and AGW is supported by the Leverhulme Trust under Research Project Grant RPG-2020-140.

References

Augenbaum, J. M. (1984). A Lagrangian method for the shallow water equations based on a Voronoi Mesh—one dimensional results. *Journal of Computational Physics*, *53*(2), 240–265. https://doi.org/10.1016/0021-9991(84)90040-8

Barlier, F., Berger, C., Falin, J. L., Kockarts, G., & Thuillier, G. (1978). A thermospheric model based on satellite drag data. *Annales Geophysicae*, *34*, 9–24.

Barnston, A. G., Mason, S. J., Goddard, L., Dewitt, D. G., & Zebiak, S. E. (2003). Multimodel ensembling in seasonal climate forecasting at IRI. *Bulletin of the American Meteorological Society*, *84*(12), 1783–1796. https://doi.org/10.1175/BAMS-84-12-1783

Bilitza, D., Altadill, D., Truhlik, V., Shubin, V., Galkin, I., Reinisch, B., & Huang, X. (2017). International Reference Ionosphere 2016: From ionospheric climate to real-time weather predictions. *Space Weather*, *15*(2), 418–429. https://doi.org/10.1002/2016SW001593

Bishop, C. H., & Abramowitz, G. (2013). Climate model dependence and the replicate Earth paradigm. *Climate Dynamics*, *41*(3–4), 885–900. https://doi.org/10.1007/s00382-012-1610-y

Bott, A. (1989). A positive definite advection scheme obtained by nonlinear renormalization of the advective fluxes. *Monthly Weather Review*, *117*(5), 1006–1015. https://doi.org/10.1175/1520-0493(1989)117<1006:apdaso>2.0.co;2

Bowman, B. R., Tobiska, W. K., Marcos, F. A., Huang, C. Y., Lin, C. S., & Burke, W. J. (2008). A new empirical thermospheric density model JB2008 using new solar and geomagnetic indices.

Bro, R., & De Jong, S. (1997). A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics*, *11*(5), 393–401. https://doi.org/10.1002/(SICI)1099-128X(199709/10)11:5⟨393::AID-CEM483⟩3.0.CO;2-L

Bruinsma, S. (2015). The DTM-2013 thermosphere model. *Journal of Space Weather and Space Climate*, *5*, A1. https://doi.org/10.1051/swsc/2015001

Bruinsma, S. L., Tamagnan, D., & Biancale, R. (2004). Atmospheric densities derived from CHAMP/STAR accelerometer observations. *Planetary and Space Science*, *52*(4), 297–312. https://doi.org/10.1016/j.pss.2003.11.004

Codrescu, M., Negrea, C., Fedrizzi, M., Fuller-Rowell, T., Dobin, A., Jakowski, N., et al. (2012). A real-time run of the coupled thermosphere ionosphere plasmasphere elecetrodynamics (CTIPe) model. *Space Weather*, *10*(2), S02001. https://doi.org/10.1029/2011SW000736

Elvidge, S. (2014). *On the use of multi-model ensemble techniques for ionospheric and thermospheric characterisation* Ph.D Thesis. University of Birmingham. Retrieved from http://etheses.bham.ac.uk/id/eprint/5526

Elvidge, S., Angling, M. J., & Nava, B. (2014). On the use of modified Taylor diagrams to compare ionospheric assimilation models. *Radio Science*, *49*(9), 737–745. https://doi.org/10.1002/2014RS005435

Elvidge, S., Godinez, H. C., & Angling, M. J. (2016). Improved forecasting of thermospheric densities using multi-model ensembles. *Geoscientific Model Development*, *9*(6), 2279–2292. https://doi.org/10.5194/gmd-9-2279-2016

Emmert, J. T. (2015). Thermospheric mass density: A review. *Advances in Space Research*, *56*(5), 773–824. https://doi.org/10.1016/j.asr.2015.05.038

Eshagh, M., & Najafi Alamdari, M. (2007). Perturbations in orbital elements of a low Earth orbiting satellite. *Journal of the Earth and Space Physics*, *33*(1), 1–12.

Fortescue, P., Swinerd, G., & Stark, J. (2011). *Spacecraft systems engineering* (4th ed.). Wiley.

Giorgi, F., & Mearns, L. O. (2002). Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "Reliability Ensemble Averaging" (REA) method. *Journal of Climate*, *15*(10), 1141–1158. https://doi.org/10.1175/1520-0442(2003)016⟨0883:COCOAU⟩2.0.CO;2

Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus*, *57A*(3), 219–233. https://doi.org/10.3402/tellusa.v57i3.14657

Häusler, K., Lühr, H., Hagan, M. E., Maute, A., & Roble, R. G. (2010). Comparison of CHAMP and TIME-GCM nonmigrating tidal signals in the thermospheric zonal wind. *Journal of Geophysical Research*, *115*(D1), D00I08. https://doi.org/10.1029/2009JD012394

Jacchia, L. G. (1977). *Thermospheric temperature, density, and composition: New models* (p. 375). SAO special report.

John, S. R., Kishore Kumar, K., Subrahmanyam, K. V., Manju, G., & Wu, Q. (2011). Meteor radar measurements of MLT winds near the equatorial electro jet region over Thumba (8.5 E): Comparison with TIDI observations. *Annales Geophysicae*, *29*(7), 1209–1214. https://doi.org/10.5194/angeo-29-1209-2011

Kharin, V. V., & Zwiers, F. W. (2002). Climate predicitions with multimodel ensembles. *Journal of Climate*, *15*(7), 793–799. https://doi.org/10.1175/1520-0442(2002)015<0793:cpwme>2.0.co;2

Mehta, P. M., Linares, R., & Sutton, E. K. (2018). A quasi-physical dynamic reduced order model for thermospheric mass density via Hermitian Space Dynamic Mode Decomposition. Retrieved from https://arxiv.org/pdf/1802.08901.pdf

Mehta, P. M., Walker, A. C., Sutton, E. K., & Godinez, H. C. (2017). New density estimates derived using accelerometers on board the CHAMP and GRACE satellites. *Space Weather*, *15*(4), 558–576. https://doi.org/10.1002/2016SW001562

Millward, G. H., Moffett, R. J., Quegan, S., & Fuller-Rowell, T. J. (1996). A coupled thermosphere-ionosphere-plasmasphere model. In R. W. Schunk (Ed.), *Step handbook of ionospheric models* (pp. 239–279). SCOSTEP.

Murray, S. A. (2018). The importance of ensemble techniques for operational space weather forecasting. *Space Weather*, *16*(7), 777–783. https://doi.org/10.1029/2018SW001861

Palmer, T. N., Doblas-Reyes, F. J., Hagedorn, R., Alessandri, A., Gualdi, S., Andersen, U., et al. (2004). Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bulletin of the American Meteorological Society*, *85*(6), 853–872. https://doi.org/10.1175/BAMS-85-6-853

Picone, J. M., Hedin, A. E., Drob, D. P., & Aikin, A. C. (2002). NRLMSISE-00 empirical model of the atmosphere: Statistical comparisons and scientific issues. *Journal of Geophysical Research*, *107*(A12), SIA15-1–SIA15-16. https://doi.org/10.1029/2002JA009430

Purnell, D. K. (1976). Solution of the advective equation by upstream interpolation with a cubic spline. *Monthly Weather Review*, *104*(1), 42–48. https://doi.org/10.1175/1520-0493(1976)104<0042:sotaeb>2.0.co;2

Qian, L., Burns, A. G., Emery, B. A., Foster, B., Lu, G., Maute, A., et al. (2014). The NCAR TIE-GCM: A community model of the coupled thermosphere/ionosphere system. AGU Geophysical Monograph Series. In *Modeling the ionosphere-thermosphere system* (Vol. 201, No. (73)). https://doi.org/10.1002/9781118704417.ch7

Reid, I. M., McIntosh, D. L., Murphy, D. J., & Vincent, R. A. (2018). Mesospheric radar wind comparisons at high and middle southern latitudes. *Earth Planets and Space*, *70*(1), 84. https://doi.org/10.1186/s40623-018-0861-1

Reigber, C., Luhr, H., & Schwintzer, P. (2002). CHAMP mission status. *Advances in Space Research*, *30*(2), 129–134. https://doi.org/10.1016/s0273-1177(02)00276-4

Ridley, A. J., Deng, Y., Toth, G., & Tóth, G. (2006). The global ionosphere-thermosphere model (GITM). *Journal of Atmospheric and Solar-Terrestrial Physics*, *68*(8), 839–864. https://doi.org/10.1016/j.jastp.2006.01.008

Rozante, J. R., Moreira, D. S., Godoy, R. C. M., & Fernandes, a. a. (2014). Multi-model ensemble: Technique and validation. *Geoscientific Model Development*, *7*(5), 2333–2343. https://doi.org/10.5194/gmd-7-2333-2014

Schunk, R. W., Scherliess, L., Eccles, V., Gardner, L. C., Sojka, J. J., Zhu, L., et al. (2016). Space weather forecasting with a multimodel ensemble prediction system (MEPS). *Radio Science*, *51*(7), 1157–1165. https://doi.org/10.1002/2015RS005888

Solomon, S. C., Woods, T. N., Didkovsky, L. V., Emmert, J. T., & Qian, L. (2010). Anomalously low solar extreme-ultraviolet irradiance and thermospheric density during solar minimum. *Geophysical Research Letters*, *37*(16), L16103. https://doi.org/10.1029/2010GL044468

Storz, M. F., Bowman, B. R., Branson, J. I., Casali, S. J., & Tobiska, W. K. (2005). High accuracy satellite drag model (HASDM). *Advances in Space Research*, *36*(12), 2497–2505. https://doi.org/10.1016/j.asr.2004.02.020

Sutton, E. K. (2008). Effects of solar disturbances on the thermosphere densities and winds from CHAMP and GRACE satellite accelerometer datas (Doctoral dissertation). Retrieved from https://www.proquest.com/dissertations-theses/effects-solar-disturbances-on-thermosphere/docview/304639074/se-2

Sutton, E. K. (2009). Normalized force coefficients for satellites with elongated shapes. *Journal of Spacecraft and Rockets*, *46*(1), 112–116. https://doi.org/10.2514/1.40940

Thompson, P. D. (1977). How to improve accuracy by combining independent forecasts. *Monthly Weather Review*, *105*(2), 228–229. https://doi.org/10.1175/1520-0493(1977)105<0228:htiabc>2.0.co;2

Titheridge, J. E. (1995). Winds in the ionosphere—A review. *Journal of Atmospheric and Terrestrial Physics*, *57*(14), 1681–1714. https://doi.org/10.1016/0021-9169(95)00091-F

Tobiska, W. K., Bowman, B. R., & Bouwer, S. D. (2009). Solar and geomagnetic indices for the JB2008 thermosphere density model. Retrieved from http://sol.spacenvironment.net/jb2008/pubs/JB2008_solar_geomag_indices-2.pdf

Tracton, S., & Kalnay, E. (1993). Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Weather and Forecasting*, *8*(3), 379–398. https://doi.org/10.1175/1520-0434(1993)008<0379:oepatn>2.0.co;2

Weisheimer, A., Doblas-Reyes, F. J., Palmer, T. N., Alessandri, A., Arribas, A., Déqué, M., et al. (2009). ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions - skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophysical Research Letters*, *36*(21), L21711. https://doi.org/10.1029/2009GL040896