Forward looking evidence based decision making for operational environmental modeling with an application to ensemble modeling

Hendrik L. Tolman
Senior Advisor for Advanced Modeling Systems
NOAA, National Weather Service, Office of Science and Technology Integration

October 2024

DOI: https://doi.org/10.25923/nkme-1926

Forward looking evidence based decision making for operational environmental modeling with an application to ensemble modeling

Hendrik L. Tolman [a]

[a] *Office of Science and Technology Integration, National Weather Service, National Oceanographic and Oceanic Administration, USA*

ABSTRACT: Evidence-based decision making is critical for improving operational environmental numerical models. This is particularly important because the development of such models moves more and more into a community-based open-source and open-science environment. Present evidence used for operational implementation decisions generally considers the present model performance only. This essay presents a simple model to assess impacts of different modeling strategies in the future. This model require estimates of present performance gaps and impacts of strategies on improvement rates of models. It is shown that such data are available for many operational (weather) models. An example application of this model to ensemble development strategies suggests that a focus on the development of a Unified (single model) Model Ensemble in an established development group is expected to provide better operational results than an Multi Model Ensemble (MME) development approach well within a typical 5 to 10 year strategic development period, whereas an MME of opportunity can still add skill at minimal costs if it consists of the combination of unified ensembles produced by different groups.

SIGNIFICANCE STATEMENT: A simple model is presented to predict (operational) model improvement in the future under different model development scenarios. It is shown that sufficient model accuracy and sustained improvement data are available for operational forecast models to use this model for model improvement. An example application to ensemble weather modeling suggests that operational centers should focus on building single-model ensembles, and that multi model ensembles are best used as ensembles of opportunity, combining results from different sources / centers, with the understanding that the examples used here based on simple metrics might not be sufficient to obtain authoritative assessments for this

1. Introduction

Development of operational environmental modeling systems at national centers is more and more performed by large teams. This coincides with a move of such centers to go to so called “Unified Modeling” approaches across scales and applications, as pioneered by the UK Met Office (e.g., Brown et al. 2012). Arguments for such an approach are both scientific (“one environment”) and practical (an efficient business model for development and maintenance). Such a unified approach is now finding its way into many operational centers, and is being adopted by the World Meteorological Organization (WMO, Mariotti et al. 2018). Strategic planning at the US National Oceanic and Atmospheric Administration (NOAA, Link et al. 2017; Tolman and Cortinas 2020a,b) is moving NOAA to an open-source and open-science Unified Forecast System (UFS) approach supported by the Earth Prediction Innovation Center (EPIC) (Jacobs 2021; Uccellini et al. 2022). Benefits of such as approach as achieved with NOAA’s operational wave models are documented in Alves et al. (2022).

The move of NOAA to a UFS approach has long been advocated by external reviews of NOAA’s operational modeling enterprise, with the caveat that all changes and improvements of operational modeling have to be driven by evidence[1]. As the development of operational systems moves from small groups at the operational centers to much larger teams including the private sector and academia, the formal definition of metrics becomes important to increase trust and efficiency within teams. In the UFS metrics are standardized in the Model Evaluation Tools (MET B. Brown et al. 2021) and a holistic set of metrics for coupled UFS models has been developed with input from a broad group of stakeholders[2].

Traditional, evidence used in decisions for operational implementations has been "instantaneous", that is, considering the present performance of models. An example of where this appears to have been detrimental for the sustained rate of improvement of operational models at NOAA can be found in the replacement of the Geophysical Fluid Dynamics Laboratory (GFDL) hurricane model with the Hurricane Weather Research and Forecasting (HWRF) hurricane model. NOAA chose to keep developing and upgrading the GFDL model while preparing the HWRF model for implementation, creating a "moving target" for the latter model. With this moving target, the process of replacing the model dragged out for roughly a decade, In

Corresponding author: Hendrik.Tolman@NOAA.gov

[1] e.g., see 2015 UMAC report at http://www.ncep.noaa.gov/director/ucar_reports/ucacn_20151207/UMEC_Final_Report_20151207-v14.pdf

[2] See Developmental Tested Center (DTC) UFS evaluation metrics website at https://dtcenter.org/events/2021/2021-dtc-ufs-evaluation-metrics-workshop

this time existing resources had to be split between the two models, which was likely detrimental for the development rate of both models. NOAA learned from this experience when the dynamic core of the global weather models was recently replaced (Ji and Toepfer 2016). After the new dynamic core was selected, NOAA chose to stop all development on the old core. This resulted in a better resourcing of the new core and a more rapid replacement of the old model, and tentatively, a more rapid long-term improvement rate during the corresponding transition period.

The example of the global model dynamic core replacement was forward looking rather than instantaneous as decisions were made on future *expectations* of performance, rather than on the *present* performance. The main purpose of the present essay is to present a simple model for predicting the sustained improvement rate of such operational models under different resourcing strategies. This model is presented in Section 2. Such a model can help in evidence based decision making with regard to strategies for model improvement. The remainder of the essay focuses on applying the improvement model to varies scenarios of ensemble modeling. The results presented here are intended as examples, and are not intended to be an in-depth analysis of the underlying models, as a full assessment of UME and MME strategies may require a similar assessment of more in-depth metrics.

Numerical modeling has been the foundation of weather forecasting for several decades. The seminal paper of Murphy (1993) identifies three element of the "goodness" of a forecast; consistency, quality and value. Value is created by decisions made by users of the forecast, which is driven by both the accuracy and the reliability of the forecast. Murphy observes that "*In general, . . . , forecasts must be expressed in probabilistic terms*". A probabilistic approach is achieved by ensemble forecasting, using a set of perturbed model runs for each individual forecast, as described in the report "*Completing the Forecast*" (National Research Council 2006).

At the core of an ensemble system is the underlying model. For NOAA's National Weather Service (NWS), the underlying (deterministic) global model is the Global Forecast System (GFS). This model forms the basis of the Global Ensemble Forecast System (GEFS), which typically has been based on a lower-resolution previous version of the GFS, using a perturbed ensemble of initial conditions. Whereas the high resolution GFS is more accurate at short forecast ranges, the lower resolution ensemble mean of the GEFS is more accurate at longer forecast ranges (see Section 3). A larger ensemble is then created by combining the GEFS with the US Navy and Canadian ensembles forming the North American Ensemble Forecast System (NAEFS). The NAEFS in turn is more accurate than the GEFS, and all other individual ensembles making up the NAEFS. Thus, the GFS is the underlying deterministic model, the GEFS is a is a Single or "Unified" Model Ensemble (UME) solely based on the GFS model, and the NAEFS is a Multi Model Ensemble (MME).

Other ensembles are designed at their core as an MME. An example of that at the NWS is the Short Range Ensemble System (SREF), which is a limited area mesoscale ensemble using various underlying mesoscale models and physics packages to build an ensemble (e.g., Du et al. 2004; Zhou and Du 2010).

Considering this, the use of MMEs is prevalent at the NWS as they tend to provide the most accurate forecast tools. At the same time, MMEs represent a business model where resources for development and maintenance of the underlying models need to be divided which is bound to have a negative impact on the improvement rates of the individual models as well as the ensembles. After performance and improvement rates of ensembles under various scenarios have been estimated in Section 3, they are applied to the simple model improvement model in Section 4. A discussion of the results is presented in Section 5 and conclusions are provided in Sections 6.

The application of the model improvement model to MME strategies is mostly intended to illustrate the potential use of this model, and does not claim to be a complete assessment of MME strategies. Nevertheless, this initial MME strategy assessment suggests that a UME strategy is to be preferred for the NWS global models, and might be feasible for regional models soon enough to focus on UMEs for such regional models. The essay focuses on NWS ensembles only to illustrate the power of forward looking assessment of model accuracy. By no means, this is intended to suggest that the NWS ensemble are the only or even the best ensemble systems in the world.

2. Modeling model improvement

Model improvement is objectively assessed using measurable metrics. A simple model for describing the evolution in time t of the value of an arbitrary metric $m(t)$, with an initial value $m_0 = m(0)$ and ideal target value $m(\infty) = m_t$, assuming a constant improvement rate in time α, can be described as

$$\frac{d}{dt}\left[m(t) - m_t\right] = -\alpha\left[m(t) - m_t\right] \quad , \tag{1}$$

which results in the simple e-folding equation

$$m(t) = m_t + (m_0 - m_t)e^{-\alpha t} \quad . \tag{2}$$

A model with a poorer initial metric m_0 but a larger improvement rate α will eventually become better than the model it is compared to. Using the suffixes 1 and 2 for the two models, the critical time t_c where the performance (value of the metrics considered) for both models is equal becomes

$$t_c = \frac{\ln(\beta+1)}{\Delta\alpha} \ , \ \beta = \frac{m_{0,2} - m_t}{m_{0,1} - m_t} - 1 \ , \ \Delta\alpha = \alpha_2 - \alpha_1 \ , \ (3)$$

where β is the relative performance gap for the models with respect to metric m, and $\Delta\alpha$ is the acceleration in model improvement rate between the two models. If model 2 is initially less accurate than model 1 ($\beta > 0$), any accelerated improvement rate $\Delta\alpha > 0$ will result in a solution for t_c. The only restriction for applying this simple model is that the metric chosen has to be bounded, that is, it has to be possible to define or estimate m_t

3. Model improvement rates

Operational computer models supporting weather forecasting have been run for many decades. Continuous monitoring of such models provides data to be used to estimate β, $\Delta\alpha$, and hence t_c in Eq. (3). Data from NOAA's global models, hurricane and convection allowing models are assessed here for the purpose of obtaining realistic improvement rates of numerical models and ensembles.

Note that for arbitrary metrics, the target value m_t typically has a theoretical value, but also a practical value. The latter occurs as measurements are of finite accuracy, so that a "zero-error" can never be actually measured. The distinction between ideal and practical error metrics will generally be ignored below, buy is illustrated where appropriate.

a. Global models

Performance data for NOAA's operational global (GFS, GEFS, NAEFS) models have been provided by the Environmental Modeling Center (EMC). Model performance is measured by the 500 hPa height anomaly correlations (ac), where $m_t = 1$, and $m(t) < 1$. EMC provided these data for 2008 through 2017 for each forecast day up to day 16 for the three modeling systems. Values for the ensembles are obtained from the ensemble mean forecast. Equation (2) was fit objectively to these data by linear regression of the logarithm of the anomaly correlation for each model, forecast day, and calendar year individually. Results are presented in Table 1.

The top part of Table 1 shows the anomaly correlations (ac) for the three models for selected forecast days, averaged over the last five years of the data set. This period is long enough to average yearly variations, and short enough to be representative for the present state of the models. As expected, the ac drops off with increasing forecast time, and the models no longer provide a skillful forecast (generally defined as $ac < 0.6$) in the 8 to 10 day forecast range. The GEFS (ensemble mean) is more accurate than the deterministic GFS, and the NAEFS (MME) is more accurate than the GEFS (UME).

The middle part of Table 1 shows the relative accuracy of the models expressed as the performance gap β of Eq. (2) (average for the last five years). For short forecast ranges, the GFS outperforms the GEFS ($\beta < 0$) as the benefits of the higher resolution of the GFS outweigh the benefits of the averaging of random errors in the GEFS. For forecast ranges larger than 6 days the GEFS is systematically over 30% more accurate that the GFS as measured with β. The NAEFS (MME) is systematically more accurate than the GEFS (UME), more so for short forecast ranges, and systematically by approximately 7% to 10% for longer forecast ranges.

Finally, the bottom part of Table 1 shows the average annual improvement rate α, obtained from curve fitting to the entire 10 year data set. For forecast ranges less than 5 days, annual improvement rates are better than 5%. However, as the ac for these ranges is close to ideal, the practical relevance of these improvement rates is limited. For forecast ranges where the ac starts showing model deficiencies, but still has predictability (6 to 10h forecast ranges) annual model improvement rates are typically 2 - 4%. Improvement of the GFS are solely due to improvements of the model (and data assimilation) whereas the improvements of the GEFS compound these improvements with improvements in ensemble techniques, and are therefore expected to be larger.

The analysis of global model ensembles at NCEP shows that the MME approach is 7-10% more accurate with respect to the ac in extended forecast ranges where the ensemble has (borderline) predictive skills and that in this range present annual improvement rates of the UME GEFS ensemble shows annual improvement rates α of 1.5 - 3%, and the MME NAEFS has annual improvement rates of 2.5 - 4%

b. Hurricane models

Traditional error metrics used for hurricane models and forecasts are the track error and the intensity error in terms of maximum wind speed. Ideal metric values for both are $m_t = 0$. The evolution of these metrics over time since 1970 is documented on the website of the National Hurricane Center (NHC)[3]. Sections 5 and 6 of this web site show the trends of the official forecast error and of the model errors, respectively. A caparison of forecast and model track errors in the two sections indicate that the forecast errors are strongly correlated to the model errors. As the forecast errors show better defined trends than the errors of individual models, the former will be used here as a proxy for the latter.

Table 2 presents annual improvement rates α in percent for the track error from the NHC data base, for both North Atlantic and Eastern North Pacific storms. Results are presented for the entire data set, or for the last 17 years

[3] https://www.nhc.noaa.gov/verification

forecast at day :	2	4	6	8	10	12	14
				ac			
GFS :	0.990	0.938	0.801	0.605	0.421	0.281	0.191
GEFS :	0.990	0.945	0.843	0.704	0.568	0.456	0.369
NAEFS :	0.992	0.952	0.858	0.726	0.596	0.489	0.407
				β (%)			
GFS → GEFS :	-3.4	13.1	27.1	34.0	34.4	33.8	28.6
GEFS → NAEFS :	25.5	14.8	10.5	8.2	7.0	6.6	6.8
GFS → NAEFS :	21.0	29.9	40.4	45.0	41.1	41.3	39.6
				α (%)			
GFS :	6.4	4.3	2.6	1.5	0.9	0.4	0.2
GEFS :	7.6	5.1	3.3	2.1	1.4	0.9	0.6
NAEFS :	9.2	6.4	4.3	3.0	2.2	1.7	1.2

TABLE 1. Global 500 hPa height anomaly correlation (ac, average for 2013-2017), performance gap (β, average for 2013-2017), and annual improvement rates (α, average for 2008-2017).

for which the extended range forecast were made (96 and 120h forecasts). Improvement rates were obtained by objectively fitting $e^{-\alpha t}$ to the data set, consistent with Eq. (2). Note that improvement rates α computed in this way are systematically larger that those obtained with a traditional linear regression as presented in other studies.

For the shortest forecast range (12h) the track errors are still close to the analysis (0h forecast) errors (see NHC web site), and hence the assumption that $m_t \approx 0$ is violated and will result in an underestimation of α. For longer forecast ranges and for the entire set of years for which data are available, annual improvement rates α are in the 3 - 4 % range, and for the last 17 years in the 4 -5.5 % range, Improvements rates in the Pacific are systematically higher than improvement rates in the Atlantic.

The NHC data base shows intensity errors for the official forecast, but not for models. It is well known that skill of intensity forecasting is more challenging to obtain than skill for the track forecast. In fact, for many years the intensity errors of the official forecast have shown much variability but limited improvements, where physical hurricane models effectively showed no skill in predicting intensity. To remedy this, the Hurricane Forecast Improvement Project (HFIP) aimed to reduce the intensity error of the HWRF model by 10% annually, with a goal of a 50% error reduction in 5 years. The focused research funded by HFIP resulted in approximately 10% annual reduction of the intensity error over 5 years as is documented in Fig. 8 of HFIP (2017) and in Tallapragada (2016).

The HFIP project shows two things. First, focused efforts on model improvements can dramatically accelerate the improvement of models. Second, stretch goals of 10% annual model improvement may be realistic with the proper focus of research and resources.

c. Regional models

To represent regional convection allowing models here, we will consider model improvement rates of the Rapid Update (RAP) and High Resolution Rapid Refresh (HRRR) models of NOAA. Model performance data for these models were provided by Curtis Alexander (personal communication). For the period 2010-2015, rms errors of temperature, humidity and wind profiles have reduced at a rate of α = 5% annually, whereas precipitation errors and biases have improved by 10 and 12% annually, respectively. Considering that the focus of the development of these models in on severe weather (precipitation), these data indicate that a focus on development in these models can result in an acceleration of improvement of 5-7% (difference in improvement rate of focus parameters versus general model behavior).

4. Prediction ensembles improvements

Equation 3 defines the critical time t_c needed for a model that is less accurate by β to catch up given a differential improvement rate $\Delta\alpha$. Figure 1 presents the corresponding lines of constant critical time t_c for given β and $\Delta\alpha$, as well as a representation of data gathered in the previous section. Figure 2 represents the same data using lines of constant differential model improvement $\Delta\alpha$ for given β and t_c. In both Figures, the red line represents t_c = 10y, loosely representative for a typical period of strategic planning.

The global ensembles discussed in Section 3.a target outlooks beyond 4-5 days. In this forecast range the MME is 7-10% more accurate than the UME. The GEFS and NAEFS show corresponding improvement rates α of 1.5-3% and 2.5-4%, respectively. Assuming that the the accelerated model improvement rate increases α by 33 to 100%, $\Delta\alpha$ is estimated as 0.5-4%. The corresponding area is identified with a green ellipse marked with 'G' in Figs. 1 and 2.

forecast :	12h	24h	36h	48h	72h	96h	120h
NA 1970 - 2016 :	2.1	2.7		3.3	3.4		
ENP 1989 - 2016 :	2.7	3.6	4.2	4.4	4.4		
NA 2001 - 2016 :	2.9	3.8	4.2	3.9	3.9	3.2	2.6
ENP 2001 - 2016 :	4.0	5.1	5.6	5.6	5.4	5.2	4.9

TABLE 2. Annual improvement rates α in percent for the hurricane track error from the official forecast of the National Hurricane Centerfor North Atlantic (NA) and Eastern North Pacific (ENP) storms.

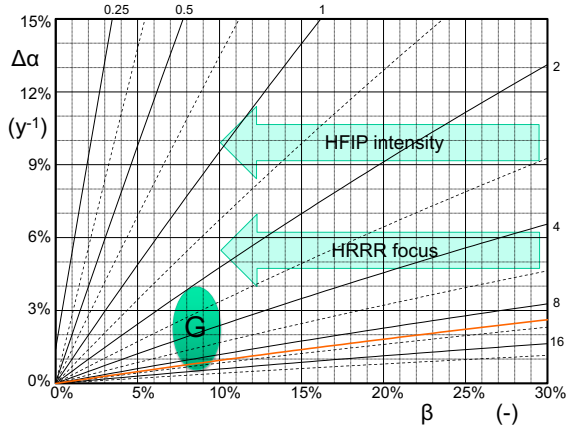

FIG. 1. Lines of constant critical time t_c in years as a function of the model performance gap β and differential model improvement rate $\Delta\alpha$. Red line corresponds to t_c = 10y. Shaded area with 'G' represents data from NWS global models.

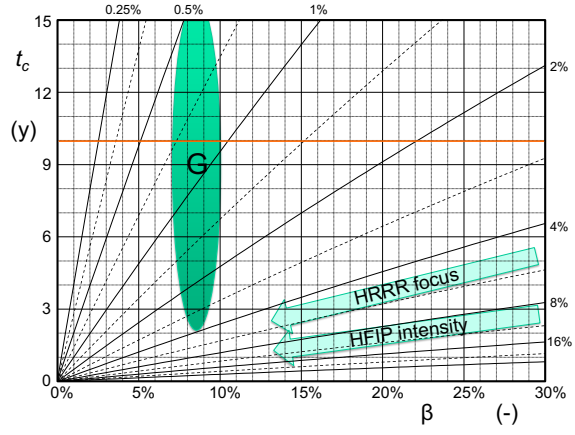

FIG. 2. Like Fig. 1 with lines of constant differential model improvement rate $\Delta\alpha$ as a function of the model performance gap β and the critical time t_c.

Hurricane intensity prediction (Section 3.b, HWRF, HFIP) and severe weather prediction with convection allowing mesoscale models (Section c, RAP and HRRR), indicate that model improvement rates can be accelerated by focused research by as much as $\Delta\alpha \approx 10\%$ and 5-7%, respectively. Because no solid data for performance gaps β for UME versus MME approaches are available here, the corresponding data are presented as annotated arrows for a broad range of β in Figs. 1 and 2.

5. Discussion

This essay presents a simple model for forward-looking evidence-based decision making for strategies employed to improve operational environmental models. The model improvement model uses initial performance gaps and estimates of impacts of strategies on improvement rates to estimate at which time a focussed strategy will result in better model behavior than is expected to be obtained with non-focussed approaches. Analysis of existing data on sustained model improvement as presented in Section 3 shows that input data for the simple model presented here is available for some operational modeling systems. As an example of the potential of this simple model, it is applied to various ensemble approached in Sec. 4. These results will be discussed below.

The irony of making decisions based on projected model improvement is that it can generally not be verified objectively, as that would require executing the different strategies side-by-side. This can be partially remedied by setting targets for model improvement rates α, and by tracking these against improvement rates of MMEs of opportunity, as was done in the HFIP project with respect to deterministic intensity forecasts (as discussed in Section 3).

The performance gaps and (accelerated) improvement rates of global model ensembles as identified by the green ellipse in Fig. 1, suggests that a UME ensemble approach will overtake the accuracy of an MME generally well within a typical strategic time frame of 10y (most of the area is above red line). Using the center of the estimated range in $(\beta, \Delta\alpha)$ space, the approach reaches benefit in typically 3-4 years. Figure 2 indicates that estimated critical or catch-up times t_c can be well beyond the 10 year limit, but that such behavior is associated with accelerated model improvement rates as small as $\Delta\alpha < 1\%$. The latter accelerated improvement rate estimates are likely conservative. Considering this, the NWS strategy to focus internally on a UME approach as used in the GEFS, while leveraging external ensemble data almost for free in the NAEFS is supported by the evidence presented here.

For the hurricane intensity data (HWRF) and the mesoscale model data (HRRR), rapidly accelerated model improvement is observed, but no performance gap data is available. Even without availability of the latter data, Fig. 2 shows that even for performance gaps as large as $\beta = 30\%$, critical catch-up times t_c are only a few years. As this is well within periods at which strategic decisions are made, it appears that the development of an UME is preferred over that of an MME at individual development groups. As with the global models, the MME approach should still be considered as an ensemble of opportunity, combining data from different modeling groups.

The example application to ensemble approaches has used a simplistic descriptions of UME and MME development approaches, with some implicit assumptions that may require a more in-depth assessment of the examples considered.

First, an underlying assumption of this study is that model development is generally resource starved. Focusing resources then will structurally accelerate model improvement. It is, however, possible that concentration of resources may over-saturate resources, which will not help development of a UME, and will be detrimental for other worthy projects. The decision to move towards a more UME-based approach should therefore always be accompanied by a resource need assessment.

Second, little attention is given to how an MME is constructed. There is a distinct difference between ensembles that are designed as an MME, or an MME that is an ensemble of opportunity, i.e., the multi-model aspect is created by adding existing models effectively for free. The later approach is generally beneficial, if only due to the increase of the ensemble size compare to the ensemble size that a group can afford to run internally, as is evident here in the data presented on the global models.

Third, this study does not address scientific differences between UME and MME approaches. For differences such as clustering and application of bias corrections, reference is made to the broadly published literature (e.g., Johnson et al. 2011a,b; Hamill and Scheuerer 2018; Gallus et al. 2019).

Fourth, a full forward looking assessment of UME and MME ensemble strategies may need to look at more relevant metrics aligned with the mission for which the ensemble is used. This is particularly true for the assessment of the accuracy of global models using the *ac* only. Such metrics could be, for instance, precipitation and temperature, weather extremes and extreme weather, and will be at least to a degree dependent on the stakeholder served by the product.

Finally, whereas the examples focus on evidence-based decision making for ensemble atmospheric models, its implications are broader. Presently, evidence based decisions are based almost exclusively on "instant gratification", that is, how much better the next implementation is. For long-term sustained model improvement it is likely better to systematically address potential long-term model improvement rates as well, and to set corresponding targets to create a long-term rather than instantaneous levels of evidence.

6. Conclusions

This essay presents a simple assessment model for future model accuracy. With this model, a critical time is estimated at which a model with poorer present behavior but with more rapid improvement will "catch up" with a model that is presently more accurate. Such a model of model accuracy allows for evidence-based decisions on model development strategies, whereas presently evidence-based decisions for operational model improvements tend to consider instantaneous model behavior only. The simple model requires data on performance gaps and impacts of strategies on improvement rates. An assessment of historical data of operational models at the US National Weather Service indicates that such data are generally available. An example application of the simple model to weather model ensemble approaches shows the potential of the forward looking approach presented here. The example results indicate that a weather ensembles based on a single underlying model will result in more accurate products well before a typical 5 to10 year strategic horizon, with the caveat that the examples focus on showing the potential of forward looking evidence based decision making, and may need more detailed assessment of more refined performance metrics to be authoritative for ensemble development strategies.

Acknowledgments. The author thanks the various providers of data as identified in the essay for making their data available. The authors thanks the anonymous reviewers for their contributions to improving this essay.

Data availability statement. This essay does not use original data from the author. All data sources are identified in the essay.

References

Alves, J.-H., H. Tolman, A. Roland, A. Abdolali, F. Ardhuin, G. Mann, A. Chawla, and J. Smith, 2022: NOAA's Great Lakes wave prediction system: A successful framework for accelerating the transition of innovations to operations. *Bull. Am. Meteor. Soc.*, https://doi.org/https://doi.org/10.1175/BAMS-D-22-0094.1.

B. Brown, B., and Coauthors, 2021: The Model Evaluation Tools (MET): More than a decade of community-supported forecast verification. *Bull. Am. Meteor. Soc.*, E782–E807, https://doi.org/https://doi.org/10.1175/BAMS-D-19-0093.1.

Brown, A., S. Milton, M. Cullen, B. Golding, J. Mitchell, and A. Shelly, 2012: Unified modeling and prediction of weather and climate: A 25-year journey. *Bull. Am. Meteor. Soc.*, **93**, 1865–1877.

Du, J., and Coauthors, 2004: The NOAA/NWS/NCEP short-range ensemble forecast (SREF) system: evaluation of an initial condition

vs multi-model physics ensemble approach. *Preprints, 16th Conference on Numerical Weather Prediction, Seattle, Washington,*, Amer. Meteor. Soc., paper 21.3, 10 pp.

Gallus, W. A., J. Wolff, J. H. Gotway, M. Harrold, , and L. Blank, 2019: The impacts of using mixed physics in the community leveraged unified ensemble. Geological and atmospheric sciences publications, Iowa State University, 21 pp.

Hamill, T. M., and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, **146(12)**, 4079–4098.

HFIP, 2017: 2016 R&D activities summary; recent results and operational implementations. HFIP Technical Report HFIP2017-1.

Jacobs, N. A., 2021: Open innovation and the case for community model development. *Bull. Am. Meteor. Soc.*, E2002–E2011, https://doi.org/https://doi.org/10.1175/BAMS-D-21-0030.1.

Ji, M., and F. Toepfer, 2016: Dynamical core evaluation test report for NOAA's Next Generation Global Prediction System (NGGPS). Report, NOAA / NWS /OSTI. https://doi.org/https://doi.org/10.25923/ztzy-qn82, 64 pp.+ Appendices.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2011a: Hierarchical cluster analysis of a convection-allowing ensemble during the hazardous weather testbed 2009 spring experiment. Part I: Development of the object-oriented cluster analysis method for precipitation fields. *Mon. Wea. Rev.*, **139(12)**, 3673–3693.

Johnson, A., X. Wang, M. Xue, and F. Kong, 2011b: Hierarchical cluster analysis of a convection-allowing ensemble during the hazardous weather testbed 2009 spring experiment. Part II: Ensemble clustering over the whole experiment period. *Mon. Wea. Rev.*, **139(12)**, 3693–3710.

Link, J., H. L. Tolman, and K. Robinson, 2017: Earth systems: NOAA's strategy for unfied modeling. *Nature*, **549 (7673)**, 458.

Mariotti, A., P. M. Rutti, and M. Rixen, 2018: Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *njp Clim Atmos Sci*, **1 (4)**.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **3**, 281–293.

National Research Council, 2006: *Completing the forecast: characterizing and communicating uncertainty for better decisions using weather and climate forecasts*. The National Academy Press, 124 pp.

Tallapragada, V., 2016: Overview of the NOAA/NCEP operational Hurricane Weather Research and Forecasting (HWRF) modeling system. *Advanced numerical modeling and data assimilation techniques for tropical cyclone prediction*, U. C. Mohanty, and G. Gopolakrishnan, Eds., Springer-Netherlands, 51–106.

Tolman, H. L., and J. Cortinas, 2020a: 2017-2018 roadmap for the production suite at NCEP[4]. Report, NOAA. 35 pp. + Appendices.

Tolman, H. L., and J. Cortinas, 2020b: A strategic vision for NOAA's physical environmental modeling enterprise[5]. Report, NOAA. 9 pp.

Uccellini, L. W., R. W. Spinrad, D. M. Koch, C. N. McLean, and W. M. Lapenta, 2022: EPIC as a catalyst for NOAA's future earth prediction system. *Bull. Am. Meteor. Soc.*, E2246–E2264, https://doi.org/https://doi.org/10.1175/BAMS-D-21-0061.1.

Zhou, B., and J. Du, 2010: Fog prediction from a multimodel mesoscale ensemble prediction system. *Wea. Forecasting*, **25**, 303–322.

---

[4] Available from https://ufscommunity.org/documents/repository/

[5] Available from https://ufscommunity.org/documents/repository/