Community-wide validation of geospace model local K-index predictions to support model transition to operations

A. Glocer[1], L. Rastätter[1], M. Kuznetsova[1], A. Pulkkinen[1], H. J. Singer[6], C. Balch[6], D. Weimer[2], D. Welling[4], M. Wiltberger[3], J. Raeder[5] and R. S. Weigel[7], J. McCollough[8], S. Wing[9]

[1]NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA.

[2]Center for Space Science and Engineering Research, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA.

[3]High Altitude Observatory, National Center for Atmospheric Research, Boulder, Colorado, USA.

[4]Department of Atmospheric, Oceanic, and Space Sciences, University of Michigan, USA.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/2016SW001387

Abstract.

We present the latest result of a community-wide space weather model validation effort coordinated among the Community Coordinated Modeling Center (CCMC), NOAA Space Weather Prediction Center (SWPC), model developers, and the broader science community. Validation of geospace models is a critical activity for both building confidence in the science results produced by the models and in assessing the suitability of the models for transition to operations. Indeed, a primary motivation of this work is supporting NOAA SWPC's effort to select a model or models to be transitioned into

[5]Space Science Center & Physics Department, University of New Hampshire, USA.

[6]Space Weather Prediction Center, NOAA, Boulder, Colorado, USA.

[7]Department of Computational and Data Sciences, George Mason University, Fairfax, Virginia, USA.

[8]Air Force Research Laboratory, Kirtland AFB, NM, USA.

[9]Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland, USA

operations. Our validation efforts focus on the ability of the models to reproduce a regional index of geomagnetic disturbance, the local K-index. Our analysis includes six events representing a range of geomagnetic activity conditions and six geomagnetic observatories representing mid- and high-latitude locations. Contingency tables, skill scores, and distribution metrics are used for the quantitative analysis of model performance. We consider model performance on an event-by-event basis, aggregated over events, at specific station locations, and separated into high- and mid-latitude domains. A summary of results is presented in this report, and an online tool for detailed analysis is available at the CCMC.

1. Introduction

Forecasting geomagnetic disturbance levels on the ground is a critical step in mitigating the potentially severe impact of geomagnetically induced currents (GICs) [e.g., *Boteler et al.*, 1998; *Pirjola*, 2005; *North American Electric Reliability Corporation*, 2012; *National Research Council*, 2008]. The science community has responded with both first principles and empirical models capable of forecasting these potentially hazardous disturbances. Before such models can be transitioned into an operational setting, a comprehensive model validation effort is required to determine the model quality and capabilities for improving services. The Community Coordinated Modeling Center (CCMC), NOAA Space Weather Prediction Center (SWPC), model developers, and the broader science community have joined together to carry out this important validation effort. This report represents the latest model validation findings in support of geospace model transition to operations.

This study builds on the prior studies of geospace model validation [*Pulkkinen et al.*, 2010, 2013; *Rastätter et al.*, 2011], and in particular is a direct follow on to [*Pulkkinen et al.*, 2013]. That study focused on the ability of models to reproduce dB/dt (the variation of ground magnetic field) at specific magnetometer locations. We encourage the reader to refer to that work, as this study is a direct follow on to that effort. As the work of [*Pulkkinen et al.*, 2013] was coming to completion, work was initiated on the present study, to consider the ability of models to reproduce a local index of geomagnetic disturbance. While the magnetic field fluctations on short times, examined in the prior study, is more directly tied to GIC prediction, a local index of variability is also useful as a convenient measure of the local risk of GIC. Moreover, it is possible that a model would have more skill in predicting the scaled range of magnetic field variability over a wider window than over a relatively short-term variation.

The *Kp* index is a commonly used global measure of geomagnetic disturbances. It is a measure on a scale of 0-9 of the average level of disturbance as measured by a scaled range of delta-B at selected geomagnetic observatories For a detailed description of how *Kp* is calculated see *Rostoker* [1972]. Local predictions of *K*, however, may differ significantly from the global *Kp*-index. The interest in predicting potential GICs and geomagnetic disturbances on a regional or local level, and the convenience of an activity index instead of a raw prediction, provides part of the motivation for this study. Additionally, we will be able to determine if the local value of the model derived *K* better represents the level of activity at a particular location than the global *Kp* index.

The layout of the paper is as follows. Section 2 describes the organization of the validation effort, Section 3 presents the metrics used to measure the model performance and Section 4 details the models. Validation results are described in Section 5, and Section 6 discusses the findings.

2. Validation setting

As noted in the previous section, the present work builds on the validation study presented by *Pulkkinen et al.* [2013]. To avoid repeating the very complete description of the validation setting provided previously, we will only provide an overview here as well as new features particular to the current study.

Six events were chosen for the study consisting of the four events from the earlier GEM Challenges [*Pulkkinen et al.*, 2010, 2013; *Rastätter et al.*, 2011] as well as two "surprise events" chosen after the modelers delivered their models to CCMC for evaluation. CCMC and NOAA SWPC scientists together choose the these two surprise events. The event list is given in Table 1.

Three high-latitude (PBQ/SNK, ABK and YKC) and mid-latitude (WNG, NEW, OTT) locations were selected. Table 2 and Fig. 1 show the locations of these stations. In the case of the global MHD models, the magnetic field variations at each magnetometer location were computed by a Biot-Savart integral over the entire domain. The integration includes all currents in the magnetosphere, as well as the field-aligned currents in the gap region between the MHD model's inner boundary and the ionosphere, and the high-latitude ionospheric currents. The CCMC tool used for the integration is described in detail by *Rastätter et al.* [2013] and is applied to each of the Global MHD models used in the study. The two empirical models (see Table 4) directly give the magnetic field at the coordinates of the station. All model runs and ground magnetic field calculations (with the exception of WingKp) were carried out at CCMC.

For every event under consideration (see Table 1), we evaluate the performance of the model by comparing the observed vs predicted local K-values at the specific magnetometer locations listed above. Throughout the paper K is calculated in the following way. First we find the maximum "Range" of ΔB in the two horizontal directions.

$$Range = \max\left[(\Delta B_{x,max} - \Delta B_{x,min}), (\Delta B_{y,max} - \Delta B_{y,min})\right] \quad (1)$$

over a three-hour window sliding by 15 minutes, where $B_{x,max}$,$B_{x,min}$,$B_{y,max}$, and $B_{y,min}$ indicate the max and min values in the window of the two horizontal components of the magnetic field (North and East in geomagnetic dipole coordinates). Strictly speaking, the quiet day variation should be subtracted before the range is calculated. However, neglecting this only introduces a relatively small error when geomagnetic activity is disturbed The Range is then divided by a station specific scaling factor. Scaling factors for stations used in this validation study are specified by IAGA through International Service of Geomagnetic Indices (ISGI) and is, generally speaking, a function of geomagnetic latitude. Those values are given in Table 2. K is then found from the scaled range using a lookup table given in Table 3. The same approach was used for both models and observations. As stated before, we follow the earlier GEM Challenges and the earlier validation study using the magnetometer stations listed in Table 2 and shown in Fig. 1. Three high- as well as three mid-latitude stations (the same as for [*Pulkkinen et al.*, 2013]) were included in the present study (Table 2). Station PBQ was no longer available in late 2007 and was therefore SNK was used. We therefore use station SNK for the 5th and 6th events. We use the results from the model and observations from [*Pulkkinen et al.*, 2013] for the time series used to calculate K in this study. No new models runs or data processing was carried out to get the time series from which we calculate the local K value. An exception to this is a rerun of the 5_WEIMER empirical model to account for errors in how that model was run in the previous study. The new results from that model (refered to as 6_WEIMER here and in the online plotting tool) are used in this analysis. 6_WEIMER has the outputs correctly rotated to geomagnetic dipole coordinates whereas 5_WEIMER does not. In addition, the CCMC had run the 5_WEIMER model with the Y component of the IMF always set to zero, due to a program error in the CCMC run scripts. The model developer found the problem which was subsequently fixed by CCMC for the rerun named 6_WEIMER. The previous dB/dt study has not yet been corrected.

3. Metrics

The model validation is largely built on event-based analyses, as described in *Pulkkinen et al.* [2013], and a distribution metric that provides new insight into model performance. The event-based analysis determines where K exceeds a threshold of k_{thres} in a three-hour sliding window. We then generate a contingency table that presents the number of correct hits, false alarms, missed events and correct no events [e.g., *Lopez et al.*, 2007]. In this work the thresholds for K were chosen to roughly correspond to the moderate ($K = 6$), and severe ($K = 8$) geomagnetic storm levels as defined by the NOAA Space Weather Scales (see e.g. http://www.swpc.noaa.gov/noaa-scales-explanation). The selected thresholds are chosen with the idea that higher K values representing stronger events are of more interest for space weather applications.

The contingency tables presented in the results section contain four entries per model evaluated: The number of times the threshold crossing was accurately predicted H (hits), the number of false predictions where a threshold crossing was predicted but not observed F, the number of observed threshold crossings missed by a model M and the number of times the model correctly predicted that no crossing occured N. These entries are used to compute the metrics used to quantify model performance. NOAA SWPC proposed three metrics for use in the final analyses: Probability of Detection (POD), Probability of False Detection (POFD) and Heidke Skill Score (HSS). For interest, we also include the Critical Success Index (CSI) as an additional skill score; however, it is not used for model ranking. For HSS, a 1 indicates a perfect score, a 0 demonstrates no skill as compared to random chance, and negative values mean that random chance has more skill than the model prediction. For POD, a 1 indicates a perfect score, while a 0 indicates that a model never makes a correct detection. For POFD, a 0 indicates a perfect score, while a 1 indicates that a model always makes false detections. For detailed descriptions of these metrics, we refer the interested reader back to the previous study by *Pulkkinen et al.* [2013].

In addition to the event tables and skill scores, we also consider a newly defined distribution metric. In this metric, we consider the distribution of model predictions when the observations are a particular value of $k = k_0$. A model that performs well in this metric would show a distribution peaked around k_0 with very little spread in the distribution. A model with significant random error would exhibit broadening of the distribution around k_0. A model with systematic error would have the distribution shifted so the peak is above or below k_0. A model with both systematic and random errors would exhibit both a shift and broadening of the distribution around k_0. In this study, we consider the distribution metric for three values of $k = 4, 6, 8$, and qualitatively compare the results to examine for the relative presence of random and systematic error in model predictions. This comparison could potentially be made more rigorous in future studies by using autocorrelation peaks.

4. Models

We include the same five models used in [*Pulkkinen et al.*, 2013]. These included empirical models by *Weimer* [2013] and *Weigel et al.* [2003] and major US global magnetohydrodynamic (MHD) models from University of Michigan [*Tóth et al.*, 2012], the Center for Integrated Space Weather Modeling (CISM) [*Wiltberger et al.*, 2004], and University of New Hampshire [*Raeder et al.*, 2008]. In addition to these models, we also include the WingKp model of Global *Kp* prediction [*Wing et al.*, 2005]. This last model was added in order to determine the "value added" of models that can predict regional K values, compared with a model currently used to predict a single global magnetic disturbance level that is assumed to apply everywhere.

As with the prior evaluation study, each model that participated in the current study was provided to CCMC. Communications with the model developers was essential to assure each model was installed correctly with correct settings and used appropriately. The WingKp model was treated differently because it is already operational at NOAA/SWPC and hence, the model was evaluated by the NOAA/SWPC staff with minimal involvement of its developer. We used the same model settings as in the previous study with final settings determined in August 2011. No model could participate if it could not run at least twice real-time on a 64 processor super computer. In otherwords, one hour of simulated time could be completed in a half hour of wall time. This is critical to ensuring models evaluated could operate in a realistic operational environment. Detailed model descriptions and milestones of model deliveries and run executions are presented in [*Pulkkinen et al.*, 2013]. All simulations, except for WingKp, were performed at CCMC using identical computational resources and were driven by ACE level 2 data for Events 2-6. As reported by [*Skoug et al.*, 2004], only low resolution data could be constructed for event 1. Additionally, the plasma density data for the event were derived from the Plasma Wave Instrument on board the Geotail Satellite.

The WingKp model was run at AFRL since it was not one of the models in the CCMC inventory. Details of this output can be found in the report by [*McCollough et al.*, 2014]. Additionally, AFRL was not able to provide results for event 3 which was outside their run window. While the other models were all driven by identical ACE level 2 data, the

WingKp model was run with the real-time ACE data, and occasionally was not able to supply a prediction due to missing data. Such predictions show up as a no data flag (K=-1) in the online plotting and are excluded from our metrics analysis. The different input data should be kept in mind when comparing model performance. WingKp was handled differently than the other models because, when available, its purpose was to compare the local prediction of K by the models under evaluation with a Kp prediction that is currently available to SWPC forecasters.

Table 4 presents some of the features of each model. Some of these models, such as the Weimer model and each of the global MHD models can be accessed through the CCMC for runs-on-request.

5. Results

All of the time series of local K values are posted online and visualizations can be made through the CCMC (http://ccmc.gsfc.nasa.gov/challenges/dBdt/metrics_results.php). Figure 2 shows an example time series of the observed vs modeled K for the event 2 (Table 1). Each model is shown in a separate panel (red line) together with the observations (black line). We chose a random mid-latitude station for this demonstration.

Event-based metrics are broken out in several different ways. First, all the events and stations are combined, as presented in Figure 3 and tables 5 and 6, to obtain an overall view of model preformance. The models are ordered from left to right by the HSS, although all the event-based skill scores, previously discussed, are presented. It is also of interest to examine the performance for different latitudes. Therefore we report the results summed over all events and high-latitude (PBQ/SNK, ABK, YKC) stations and mid-latitude (WNG, NEW, OTT) stations. Figures 4 and 5 show the performance for high-latitude stations and mid-latitude stations respectively. Other configurations were also considered such as grouping the results by the first four events that were known to the modelers ahead of the study, and the two events added later. However, in the interest of brevity the associated tables are not included here. We note that caution must be taken when determining groupings or setting thresholds to ensure that there are enough threshold crossing events. To that end we do not focus on individual magnetometer but rather the groupings specified above. The smallest number of threshold crossings in any grouping considered is 171 out of 1422 total events for midlatitude magnetometers with a threshold of 8.

As described in Section 3, we also incorporate a "distribution" metric. The concept behind this metric is as follows: We examine the distribution of model predictions at a particular station for an observed K at that same station. Although we do not employ a mathematically rigorous analysis of the model performance in the distribution metric, a great deal can still be learned by visual inspection of the distributions. For instance, a peak shifted to the left represents a systematic under-prediction while a peak shifted to the right represents a systematic over prediction. When taken in conjunction with the contingency tables and skill scores the results can be quite illuminating. A model that has a high-probability of false detection, for instance, could have those false detections as a result of a systematic error causing the model to consistently predict higher values, random errors causing the model to result in more false detections, or a combination of both. The contingency tables alone cannot pinpoint the type of error, but including the distribution metric can provide insight into the cause for, in this case, the false detection.

When evaluating results from using the distribution metric, we consider the results station-by-station to gain a more granular picture of model performance. One important factor to keep in mind is that the number of events decreases for $K = 8$ and may be very small when considering the distribution on a station-by-station basis (on the order of 50 events). To be concise, here we only present a single example of the distribution metric; however, all the figures are made available in the online supplementary material. Figure 6 shows an example of the distribution metric for the 6_WEIMER Model. The figure presents results for K=4 (left column), K=6 (middle column), or K=8 (right column). Additionally, each row presents results for a different magnetometer station. In the following paragraphs we will summarize the results of this distribution metric for each model, starting with the 6_WEIMER and 9_SWMF models which where the top performers in the event-based metrics.

For both mid-latitude stations (OTT and NEW), for observed K=4 and K=8, the distribution of model predictions for the 6_WEIMER Model is peaked below the observations. For K=6 the distribution of model predictions is peaked right at 6 for the mid-latitude stations. For high-latitude stations for all observed values of K the distribution is seen to be shifted to the left representing a systematic under prediction. This pattern seems consistent with the event based studies when the model showed low POFD (apparently due to the systematic under-prediction) and the strongest performance among models for mid-latitude stations when the K threshold is set to 6, but worse performance for higher K threshold and high-latitude.

The 9_SWMF Model distribution results for mid-latitude stations are typically peaked at or near the correct values of K. Some moderate spread in the distributions are present indicating the presence of some random error. The same largely holds true for high-latitude results with the spreading a bit more pronounced. Also a slight systematic shift towards under-prediction is seen when the observed K=8. This is consistent with the trend seen in the event studies that performance for 9_SWMF was stronger for mid-latitude compared to high-latitude. It is also consistent with the finding from the event table that 9_SWMF has higher skill for threshold of K=8 (compared to K=6) for mid-latitude, but the reverse is true for high-latitude. Note that virtually identical results are found for 9a_SWMF, which is expected, as it is the same model run, but the magnetometer timeseries from which K is calculated is provided by the model's internal tools rather than the CCMC tool. This provides an independent check of the CCMC tool for calculating the magnetometer timeseries.

For the 2_LFM-MIX Model the distribution of model predictions for an observed K tend to peak below the observed value of K for both mid- and high-latitude stations. This shift in the peak of the distribution relative to the observed K is indicative of a systematic under-prediction by the model. The 2_LFM-MIX model was found to have extraordinarily low POFD in the event based analysis which is likely a result of this systematic shift. Some modest evidence of random error is visible in the spreading of the distribution, but it is not enough to result in significant false detections for the K thresholds considered.

The 4_OPENGGCM Model demonstrates a large number of occurrences in the model predictions of K values greater than then observed K. Sometimes this is a systematic shift in the distribution (e.g., WNG and NEW , K=4), and sometimes it appears to be more random error (e.g., OTT K=4 and NEW K=6). Regardless of whether the shift is systematic or random, the high-occurrence of predictions significantly exceeding the observations, particularly for mid-latitude stations and lower K values, results in a large rate of false detection (even if true detections are plentiful). This finding is consistent with the high-POFD and high POD exhibited by 4_OPENGGCM in the event studies.

For the 2_WEIGEL Model, for both mid- and high-latitude stations, and for all choices of observed K, the distribution of model predictions is peaked below the observations. Such a shift represents a systematic under-prediction of the model. As a result, the model is likely to have a low POFD. These findings are consistent with the event-based analysis which demonstrates that the 2_WEIGEL model has low POFD.

Finally, the WingKp Model demonstrates a very large spread indicating significant random error when trying to predict K using the global Kp prediction. For K=8, the results are more peaked at the correct value of K although some random error is still visible. The results are similar for high-latitude which is consistent with the event based analysis. However, not including the strongest storm for this model may introduce some bias in the analysis for larger K values. The results for station PBQ are particularly good with peaks at the correct values of K, albeit with some spread. However, the results for stations YKC and ABK exhibit significant random error for all values of K. As WingKp produces a single global prediction of Kp, and we are using that prediction for local K predictions, some error is to be expected. From this type of analysis we can see that the error is mostly random in nature.

In summary, the distribution metric, is quite useful in understanding and interpreting the results of the event based metrics. The distribution metric reveals the presence of systematic and random errors and how that can affect the POD and POFD (either positively or negatively).

6. Discussion

This work describes another phase of the geospace model validation effort building on the earlier GEM modeling challenges and the dB/dt validation study summarized in *Pulkkinen et al.* [2013]. The work was carried out in coordination among the CCMC, NOAA SWPC, modelers and the science community. The focus of the effort was to evaluate the ability of geospace models to predict the local K index and moreover to evaluate the potential value added of a local prediction over the global prediction.

We considered two types of metrics in evaluating the model K prediction: skills scores calculated from event-based contingency tables and a distribution metric. The skills scores (POD, POFD and HSS) from event-based contingency tables for different K thresholds were the primary metric used to rank the models. In particular, the HSS, reflects how much better a model skill is compared to random chance. The derived contingency tables were compiled by grouping all the stations and events together, by separating high-latitude stations and mid-latitude stations for all events, and by separating events into those known to the model developer ahead of time (first four events) and the surprise events selected after models were delivered to CCMC for evaluation (last two events). These different groupings allow us to draw more detailed conclusions about model performance and suitability for forecasting K values at mid-latitude vs high-latitude and for strong events vs very strong events. The distribution metric was an additional tool used to gain insight into aspects of model performance such as revealing random error and systematic errors.

In terms of actual model performance, the 9_SWMF and 9a_SWMF models were consistently strong performers in all the metrics almost always ranking near the top in all categories. The model had relatively high-POD and low POFD resulting in a HSS that was always among the best. The distribution metric revealed the presence of a moderate amount of random error and limited systematic error. We reiterate that similar performance is expected for 9_SWMF and 9a_SWMF since they are actually the same model except for how the ground-magnetic field perturbation is calculated.

The 2_LFM-MIX model typically had lower performance compared to other models as measured by the HSS. The exception was the last two events for mid-latitude where the model performance was in the middle of the pack. The model typically exhibited lower POD and POFD. The distribution metric shows a clear tendency of this model to under-predict K and that likely results in the lower POD, POFD, and HSS. We note that these results are consistent with the earlier dB/dt study in which the 2_LFM-MIX model performed worse for larger thresholds of magnetic perturbation. It is possible that the model would perform better for lower K thresholds for calculating the contingency tables, just as the model did better in the dB/dt study for lower thresholds. However, the present study is focused primarily on model ability to detect strong and very strong disturbances, not small or moderate disturbances. A cursory examination of a lower threshold of K=4 did not result in a significant change in the ordering of models by performance (although the HSS increased). Another factor contributing to the poor model performance during storm-time is the lack of of ring current model. More recent version of the LFM include coupling with the Rice Convection Model (RCM) [*Pembroke et al.*, 2012] and are likely to improve performance on these metrics.

The 6_WEIMER statistical model performed exceptionally well for mid-latitudes for a threshold of K=6, the top performer in this category. The model performance decreased significantly for mid-latitudes with a threshold of K=8, but the performance was still strong. In contrast to mid-latitudes the model performance dropped significantly at high-latitude for both K thresholds.

The 4_OPENGGCM model had mixed performance. It generally had very good POD, but it also had a consistently elevated POFD. As seen from the distribution metric results, the model had a tendency to over predict, leading to a high POD and high POFD. As a result, sometimes the model has a good HSS and sometimes worse depending on how strongly the POD outweighed the POFD. Significant random and systematic error was likely the cause of the the higher POFD. Regardless of the cause, and overall result on the HSS, an elevated POFD is a concern that needs to be considered in an operational setting. The model did perform better in the last two events compared to the first four.

The 2_WEIGEL model was never the top-performing model, but it was also never the worst performing model as measured by HSS. The distribution metric results showed that the model typically underpredicted the observations, and as a result, have an exceedingly low POFD with a reasonable POD.

One of the key questions this study addresses is: "How well do geospace models predict local geomagnetic activity (K) compared to representing that activity by the global Kp index?" To answer that question we included in our analysis the WingKp model, which is currently used by SWPC as one method for predicting short-term Kp. The WingKp model never ranked at the bottom or the top of the model rankings based on its HSS. Interestingly, the model used in this way was also often not the lowest performing model, indicating that using the WingKp prediction of global Kp (as a local K prediction) would actually exhibit higher skill than using the local K predicted by some models. However, the POFD was typically elevated compared to other models. An elevated POFD raises concerns for using the global Kp prediction from WingKp for local forecasts of K, but it also demonstrates the potential value of a local K forecast. All local K forecasts (except for 4_OPENGGCM) consistently had much lower POFD than WingKp. However, the POD score is near the top in some cases. One caution when interpreting these results is that the WingKp model used different solar wind inputs than the other models. It is possible that the results could have been somewhat different had the same input solar wind parameters been used.

One consideration for transition to operations is lead times for model prediction. The main constraint in this regard is the input data from ACE which arrives at most one hour ahead of the event. The empirical models in this study can provide a practically instantaneous prediction with very modest computing resources while the MHD models are more resource intensive. As noted earlier, one requirement for the MHD models was they could run in twice real-time on a moderately sized supercomputing cluster. If larger computational resources are available these models could run faster. Nevertheless, the empirical models will always be more computationally efficient than the MHD models.

All the models had positive HSS demonstrating better prediction skill than random chance. Moreover, we found most results consistent with the dB/dt study of *Pulkkinen et al.* [2013]. When considering all events, a POD of around 70% is found for the top performing models for mid-latitude stations, even with a K threshold of 8. For high-latitude stations, the POD possible for top performing models drops to around 50%. In either case, the POFD for most models is exceedingly low for the thresholds considered. Whether this performance is sufficient for current space weather prediction needs, or if further improvement is required is not a question addressed in this study. We also note that this study only evaluates model prediction of K and therefore cannot be used to draw conclusions about how those models would perform when predicting other quanties, even closely related ones. Indeed, it is entirely possible to that a model can produce a value of K that is very close to that determined from the measurements, while having ΔB predictions with signs that are mostly opposite of the measured value. As a result of the model evaluation conducted by CCMC in coordination with modelers and NOAA-SWPC, NOAA-SWPC has decided to transistion the SWMF model to space weather operations and to give further consideration to the Weimer model. As the models continue to improve and evolve, it is likely that more geospace models will transition to operations for purposes of addressing specific user needs, for incorporating improved models, and for ensemble modeling. Indeed, this validation is just one step on the path of operationalizing state-of-the-art codes for space weather forecasting.

Acknowledgments.

The data from the ground based magnetic observatories was critical to this study. As such, we thank the institutions that support those observatories as well asl INTERMAGNET for promoting high-standards of practice (www.intermagnet.org). The National Center for Atmospheric Research is supported by the National Science Foundation. All model output used in the analysis is available through the CCMC as described in the manuscript.

References

Boteler, D.H., R.J. Pirjola, and H. Nevanlinna (1998), The Effects of Geomagnetic Disturbances on Electrical Systems at the Earth's Surface, *Adv. Space Res.*, 22, 17-27.

Lopez, R. E., S. Hernandez, M. Wiltberger, C.-L. Huang, E. L. Kepko, H. Spence, C. C. Goodrich, and J. G. Lyon (2007), Predicting magnetopause crossings at geosynchronous orbit during the Halloween storms, *Space Weather*, 5, S01005, doi:10.1029/2006SW000222.

McCollough, J. P., S. L. Young, and W. R. Frey (2014), Real-Time Validation of the *Kp* Predictor Model, *AFRL Tech. Rep.*, AFRL-RV-PS-TR-2015-0073.

National Research Council (2008), Severe Space Weather Events-Understanding Societal and Economic Impacts: A Workshop Report, *The National Academies Press*, Washington, DC.

North American Electric Reliability Corporation GMD Task Force (2012), 2012 Special Reliability Assessment Interim Report: Effects of Geomagnetic Disturbances on the Bulk Power System, *NERC*, February 2012.

Pembroke, A., F. Toffoletto, S. Sazykin, M. Wiltberger, J. Lyon, V. Merkin, and P. Schmitt (2012), Initial results from a dynamic coupled magnetosphere-ionosphere-ring current model, *J. Geophys. Res.*, 117(A2), doi:10.1029/2011JA016979.

Pirjola, R., Effects of space weather on high-latitude ground systems, *Advances in Space Research*, 36, 2231-2240, 2005.

Pulkkinen, A., L. Rastätter, M. Kuznetsova, M. Hesse, A. Ridley, J. Raeder, H.J. Singer, and A. Chulaki (2010), Systematic evaluation of ground and geostationary magnetic field predictions generated by global magnetohydrodynamic models, *Journal of Geo-*

physical Research, 115, A03206, doi:10.1029/2009JA014537.

Pulkkinen, A., M. Kuznetsova, A. Ridley, J. Raeder, A. Vapirev, D. Weimer, R. S. Weigel, M. Wiltberger, G. Millward, L. Rastätter, M. Hesse, H. J. Singer and A. Chulaki (2013), Geospace Environment Modeling 2008-2009 Challenge: ground magnetic field perturbations, *Space Weather*, Vol. 9, S02004, doi:10.1029/2010SW000600.

Pulkkinen, A., L. Rastätter, M. Kuznetsova, H. Singer, C. Balch, D. Weimer, G. Toth, A. Ridley, T. Gombosi, M. Wiltberger, J. Raeder, R. Weigel(2011), Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations, *Space Weather*, Vol. 11, doi:10.1002/swe.20056.

Raeder, J., D. Larson, W. Li, E. L. Kepko, and T. Fuller-Rowell (2008), OpenGGCM simulations for the THEMIS mission, *Space Sci. Rev.*, 141, 535, doi:10.1007/s11,21400894215.

Rastätter, L., M. Kuznetsova, A. Vapirev, A. Ridley, M. Wiltberger, A. Pulkkinen, M. Hesse and H.J. Singer (2011), Geospace Environment Modeling 2008-2009 Challenge: geosynchronous magnetic field, Space Weather, Vol. 9, S04005, doi:10.1029/2010SW000617.

Rastätter, L., G. Toth, M. M. Kuznetsova, and A. A. Pulkkinen (2014), CalcDeltaB: An efficient postprocessing tool to calculate ground-level magnetic perturbations from global magnetosphere simulations, *Space Weather*, 11, doi:10.1002/2014SW001083

Rostoker, G. (1972), Geomagnetic Indices, *Reviews of Geophysics*, 10, 935–950, doi:10.1029/RG010i004p00935.

Skoug, R. M., Gosling, J. T., Steinberg, J. T., McComas, D. J., Smith, C. W., Ness, N. F., Hu, Q., Burlaga, L. F., (2004), Extremely high-speed solar wind: 2930 October 2003, *J. Geophys. Res.*, 109, A09102, doi:10.1029/2004JA010494.

Tóth, G. et al (2012), Adaptive numerical algorithms in space weather modeling, *J. Comput. Phys.*, 231.

Weigel, R. S., A. J. Klimas, D. Vassiliadis (2003), Solar wind coupling to and predictability of ground magnetic fields and their time derivatives, *J. Geophys. Res.*, 108 (A7), 1298, doi:10.1029/2002JA009627.

Weimer, D. R. (2013), An empirical model of ground-level geomagnetic perturbations, *Space Weather*, 11, 107-120, doi: :10.1002/swe.20030.

Wiltberger, M., W. Wang, A. G. Burns, S. C. Solomon, J. G. Lyon, and C. C. Goodrich (2004), Initial results from the coupled magnetosphere ionosphere thermosphere model: magnetospheric and ionospheric responses, *Journal of Atmospheric and Solar-Terrestrial Physics*, 66(1), 1411-1423, doi:10.1016/j.jastp.2004.03.026.

Wing, S., J. R. Johnson, J. Jen, C.-I. Meng, D. G. Sibeck, K. Bechtold, J. Freeman, K. Costello, M. Balikhin, and K. Takahashi (2005), *Kp* forecast models, *Journal of Geophysical Research*, 110, A04203, doi:10.1029/2004JA010500.

Table 1. Geospace events studied in the validation activity. The last two columns give the minimum *Dst* index and the maximum *Kp* index of the event, respectively.

Event #	Date and time	min(*Dst*)	max(*Kp*)
1	October 29, 2003 06:00 UT - October 30, 06:00 UT	-353 nT	9
2	December 14, 2006 12:00 UT - December 16, 00:00 UT	-139 nT	8
3	August 31, 2001 00:00 UT - September 1, 00:00 UT	-40 nT	4
4	August 31, 2005 10:00 UT - September 1, 12:00 UT	-131 nT	7
5	April 5, 2010 00:00 UT - April 6, 00:00 UT	-73 nT	8-
6	August 5, 2011 09:00 UT - Aug 6, 09:00 UT	-113 nT	8-

Table 2. The locations of the geomagnetic observatories used in the study.

Station name	Station code	Geomagnetic lat	Geomagnetic lon	Scaling Factor
Yellowknife	**YKC**	68.9	299.4	3.0
Newport	**NEW**	54.9	304.7	1.4
Poste-de-la-Baleine	**PBQ**	65.5	351.8	3.0
Sanikiluaq	**SNK**	66.4	356.1	3.0
Ottawa	**OTT**	55.6	355.3	1.5
Abisko	**ABK**	66.1	114.7	3.0
Wingst	**WNG**	54.1	95.0	1.0

Table 3. Look up table to determine K from scaled range of ΔB.

K-index	nT range
0	$0 \leq$ Range of $\Delta B < 5$
1	$5 \leq$ Range of $\Delta B < 10$
2	$10 \leq$ Range of $\Delta B < 20$
3	$20 \leq$ Range of $\Delta B < 40$
4	$40 \leq$ Range of $\Delta B < 70$
5	$70 \leq$ Range of $\Delta B < 120$
6	$120 \leq$ Range of $\Delta B < 200$
7	$200 \leq$ Range of $\Delta B < 330$
8	$330 \leq$ Range of $\Delta B < 500$
9	$500 \leq$ Range of ΔB

Table 4. Models analyzed in the validation effort. Each model is assigned a unique model identifier given by the leftmost column of the table. The table indicates the model description, and if applicable, the number of cells and the minimum spatial resolution used in the global MHD part of the model. See text in Section 4 for details.

Identifier	(model version) Model	Grid (# of cells, min. res.)
2_LFM-MIX	(LTR-2.1.1) LFM coupled with ionospheric electrodynamics	163,000, 0.4 R_E
3_WEIGEL	empirical model	N/A
4_OPENGGCM	(OpenGGCM 4.0) global MHD coupled with CTIM	3.9 million, 0.25 R_E
6_WEIMER	empirical model	N/A
9_SWMF	(SWMF 2011-01-31) BATS-R-US coupled with RIM and RCM	1 million, 0.25 R_E
9a_SWMF	Same as 9_SWMF but using internal SWMF calculation for magnetometer timeseries	
Acronyms:		
RIM	Ridley Ionosphere Model	
RCM	Rice Convection Model	
CTIM	Coupled Thermosphere Ionosphere Model	

Table 5. Table for all stations, threshold 6

Run	n_event	n_noevent	H	F	M	N	HSS	CSI	POD	POFD
9_SWMF	1240	1532	801	74	439	1458	0.61	0.61	0.65	0.05
9a_SWMF	1240	1532	752	38	488	1494	0.60	0.59	0.61	0.02
6_WEIMER	1240	1532	605	20	635	1512	0.50	0.48	0.49	0.01
2_WEIGEL	1240	1532	537	25	703	1507	0.44	0.42	0.43	0.02
WingKp	1151	1117	722	279	429	838	0.38	0.50	0.63	0.25
4_OPENGGC	1240	1532	803	425	437	1107	0.37	0.48	0.65	0.28
2_LFM-MIX	1240	1532	353	26	887	1506	0.29	0.28	0.28	0.02

Table 6. Table for all stations, threshold 8

Run	n_event	n_noevent	H	F	M	N	HSS	CSI	POD	POFD
9a_SWMF	395	2377	201	55	194	2322	0.57	0.45	0.51	0.02
9_SWMF	395	2377	210	80	185	2297	0.56	0.44	0.53	0.03
2_WEIGEL	395	2377	116	41	279	2336	0.37	0.27	0.29	0.02
4_OPENGGC	395	2377	139	145	256	2232	0.33	0.26	0.35	0.06
WingKp	370	1898	121	137	249	1761	0.29	0.24	0.33	0.07
6_WEIMER	395	2377	79	18	316	2359	0.28	0.19	0.20	0.01
2_LFM-MIX	395	2377	42	11	353	2366	0.16	0.10	0.11	0.00

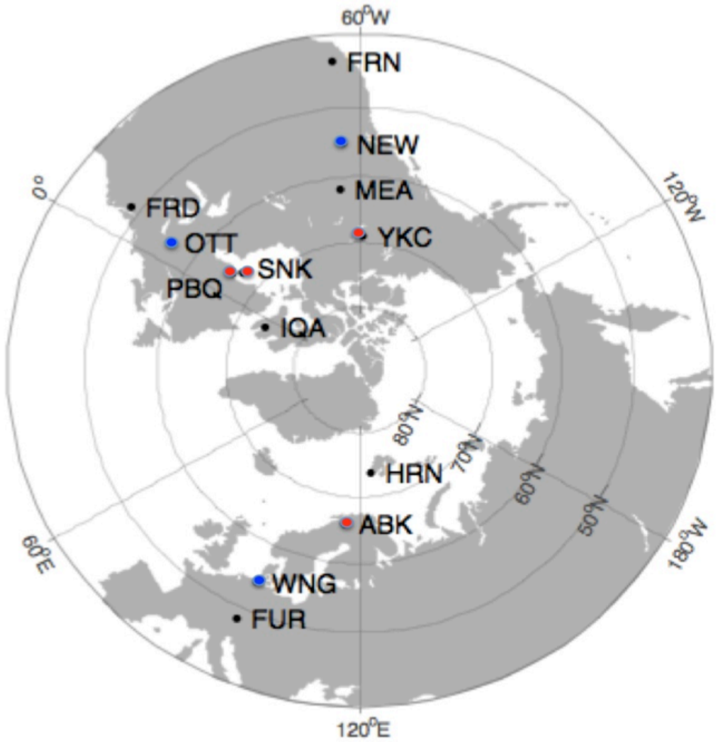

Figure 1. The locations and the station codes of the geomagnetic observatories used in the study. Geomagnetic dipole coordinates are used. Red and blue circles indicate high-latitude and mid-latitude stations, respectively, used in the final analyses in Section 5.

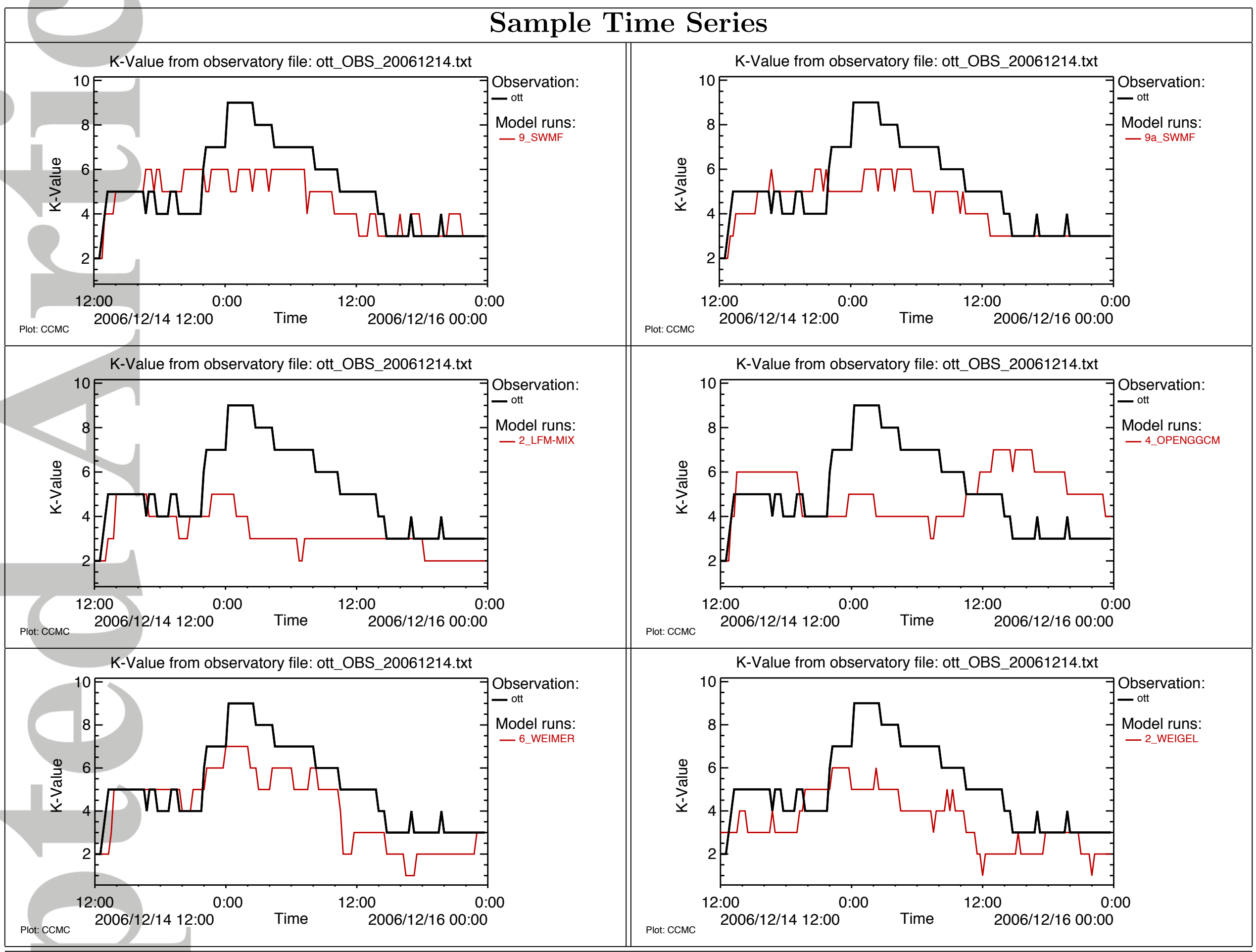

Figure 2. Time series of the observed (Black) and modeled (Red) K predictions for a particular mid-latitude station (OTT). Each panel shows a different model's prediction.

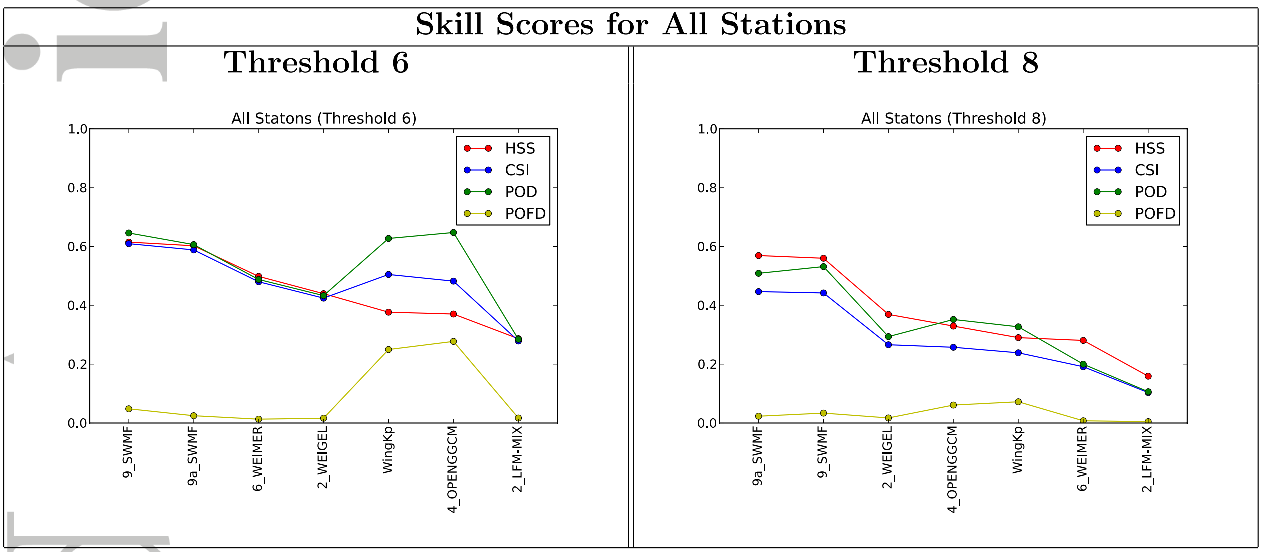

Figure 3. Heidke Skill Score (HSS), Critical Success Index (CSI), Probability of Detection (POD) (blue curve) and Probability of False Detection (POFD) (yellow curve) defined in Section 3 for the K thresholds 6 (left panel) and 8 (right panel). POD and POFD obtained by integrating over the three mid-latitude stations and the three high-latitude stations. The models (see Table 4) are ordered according to their HSS. The model with the largest HSS is the leftmost in all panels.

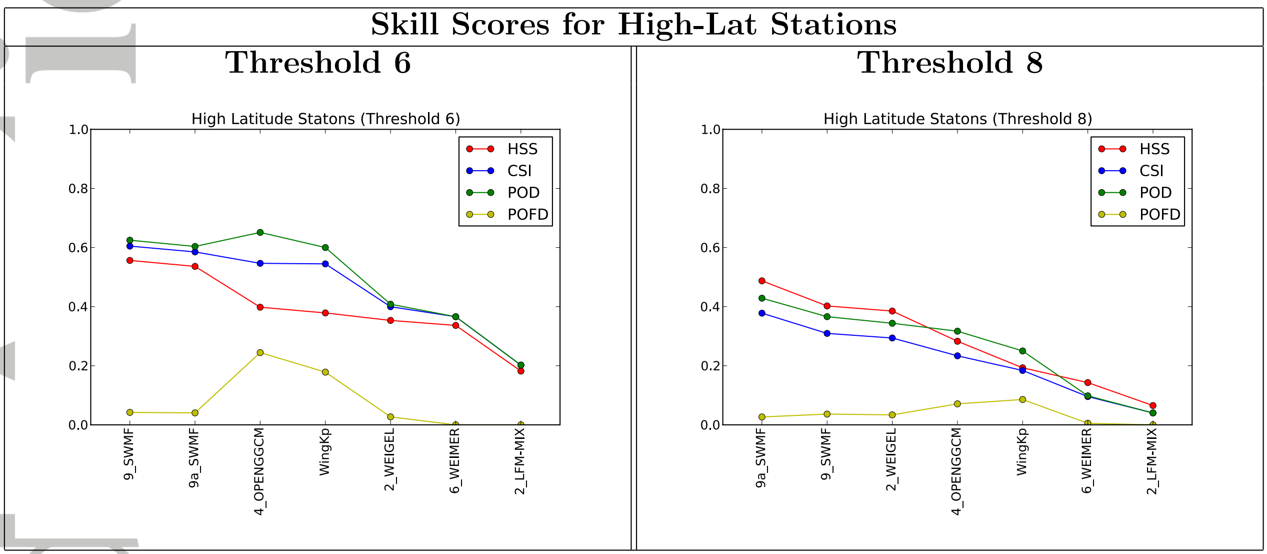

Figure 4. Heidke Skill Score (HSS) (red curve), Critical Success Index (CSI) (blue curve), Probability of Detection (POD) (green curve) and Probability of False Detection (POFD) (yellow curve) defined in Section 3 for the K thresholds 6 (left panel) and 8 (right panel). POD and POFD are obtained by integrating over the three high-latitude stations. The models (see Table 4) are ordered according to their HSS. The model with the largest HSS is the leftmost in all panels.

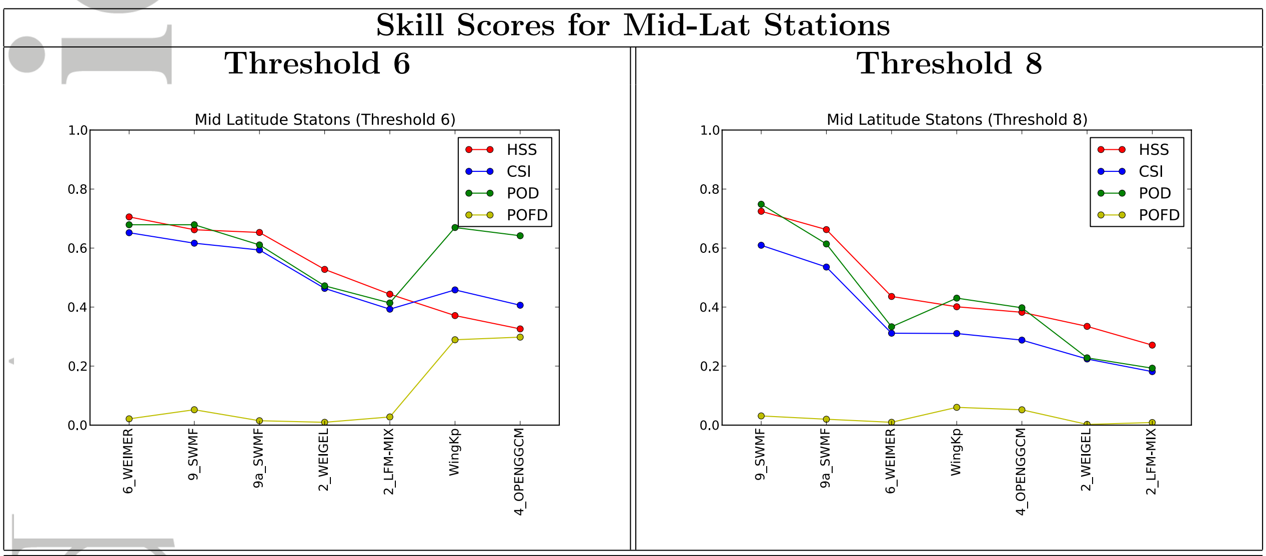

Figure 5. Heidke Skill Score (HSS) (red curve), Critical Success Index (CSI) (blue curve), Probability of Detection (POD) (green curve) and Probability of False Detection (POFD) (yellow curve) defined in Section 3 for the K thresholds 6 (left panel) and 8 (right panel). POD and POFD are obtained by integrating over the three mid-latitude stations. The models (see Table 4) are ordered according to their HSS. The model with the largest HSS is the leftmost in all panels.

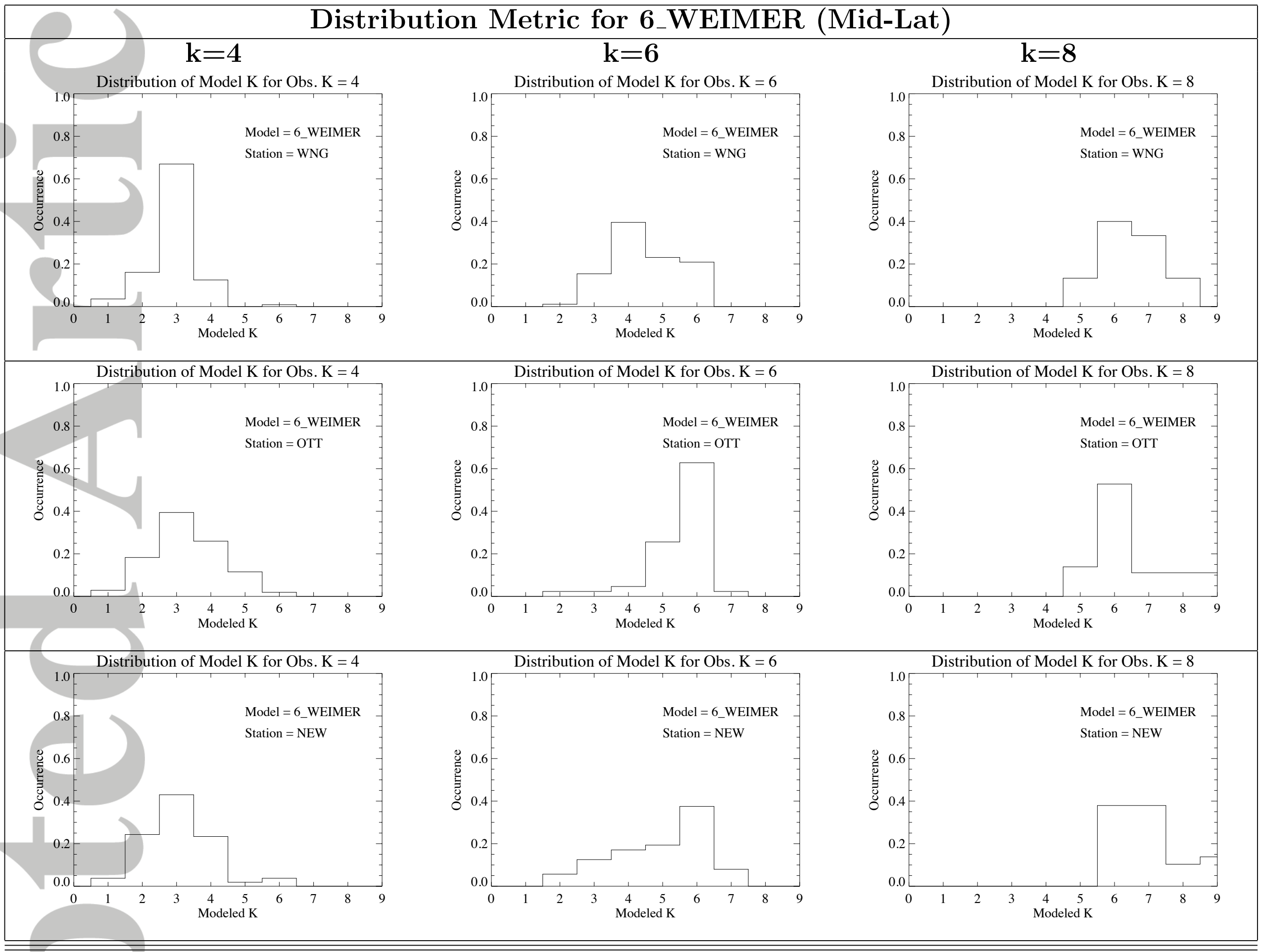

Figure 6. Distribution of 6_WEIMER Model predictions when K=4 (left column), K=6 (middle column), and K=8 (right column). Each row presents results for a different mid-latitude station.