Towards GIC forecasting: Statistical downscaling of the geomagnetic field to improve geoelectric field forecasts

C. Haines [1], M.J. Owens[1], L. Barnard[1], M. Lockwood[1], C.D. Beggan [2], A.W.P. Thomson [2], N.C. Rogers [3]

[1] Department of Meteorology, University of Reading, Earley Gate, PO Box 243, Reading, RG6 6BB, UK
[2] British Geological Survey, Research Ave South, Riccarton, Edinburgh, EH14 4AP, UK
[3] Department of Physics, Lancaster University, Lancaster LA1 4YB, UK

Key Points:

- Operational global MHD models do not fully capture the ground-level magnetic field variability important for modelling induction hazards
- We provide a proof of concept model to statistically introduce realistic, high-resolution perturbations with which to drive an impacts model
- Our downscaling scheme outperforms a reference linear-interpolation approach under a range of metrics

Corresponding author: Carl Haines, carl.haines@pgr.reading.ac.uk

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: 10.1029/2021SW002903.

Abstract

Geomagnetically induced currents (GICs) are an impact of space weather that can occur during periods of enhanced geomagnetic activity. GICs can enter into electrical power grids through earthed conductors, potentially causing network collapse through voltage instability or damaging transformers. It would be beneficial to power grid operators to have a forecast of GICs that could inform decision making on mitigating action. Long lead-time GIC forecasting requires magnetospheric models as drivers of geoelectric field models. However, estimation of the geoelectric field is sensitive to high-frequency geomagnetic field variations which operational global magneto-hydrodynamic models do not fully capture. Furthermore, an assessment of GIC forecast uncertainty would require a large ensemble of magnetospheric runs, which is computationally expensive. One solution that is widely used in climate science is "downscaling", wherein sub-grid variations are added to model outputs on a statistical basis. We present proof-of-concept results for a method that temporally downscales low-resolution magnetic field data on a 1-hour timescale to 1-minute resolution, with the hope of improving subsequent geoelectric field magnitude estimates. An analogue ensemble (AnEn) approach is used to select similar hourly averages in a historical dataset, from which we separate the high-resolution perturbations to add to the hourly average values. We find that AnEn outperforms the benchmark linear-interpolation approach in its ability to accurately drive an impacts model, suggesting GIC forecasting would be improved. We evaluated the ability of AnEn to predict extreme events using the FSS, HSS, cost/loss analysis and BSS, finding that AnEn outperforms the "do-nothing" approach.

Plain Language Summary

Forecasting space weather impacts on ground-based systems, such as power grids, requires the use of computer simulations of the disturbance of the Earth's magnetic field by the solar wind. However, these computer simulations are often too smooth, underestimating small and fast variations in the Earth's magnetic field which are important for modelling induction hazards that may affect power grids. In this paper we present a proof-of-concept scheme that attempts to introduce realistic high-frequency variations using the idea of looking at how the field has previously behaved in historical events. We test the model and find that it allows for better impact forecasting than if our scheme is not used.

1 Introduction

Intensification of magnetospheric and ionospheric current systems drives changes in the geomagnetic field measured on the ground ($\frac{d\mathbf{B}}{dt}$) which induces an enhanced geoelectric field, as expressed by the Maxwell-Faraday equation. The induced geoelectric field drives currents within the Earth that can enter grounded conducting networks as geomagnetically induced currents (GICs) (Koskinen et al., 2017; Pulkkinen et al., 2017). GICs can flow into the power grid through earthing points at substations (Oughton et al., 2017; Cannon et al., 2013), particularly in regions with high ground resistance, as the geoelectric field is larger and the network provides a more favourable path for GICs to flow. The quasi-DC signal introduced into an AC grid system can lead to half cycle saturation in transformers causing degradation and, in extreme cases, destruction, failure and system collapse. The geomagnetic field can be used as a proxy for potential ground effects and GIC studies commonly use the time derivative $\frac{d\mathbf{B}}{dt}$ to quantify potential effects.

Nowcasting and advanced forecasting of geomagnetic disturbances is generally achieved through global magnetohydrodynamic (MHD) models (Welling, 2019), driven with near-Earth solar wind observations or, for increased lead time, the output of solar-wind simulations (Merkin et al., 2007). The ground-level magnetic field, which is typically extrap-

olated from much higher in the magnetospheric domain, is used to drive geoelectric field models. Empirical models also exist (Weimer, 2013, 2019).

An example of global MHD system is the Space Weather Modeling Framework (SWMF Tóth et al., 2005, 2012). Other widely used MHD models include the Lyon-Fedder-Mobarry (LFM) model (Lyon et al., 2004) and the Open Global Generalized Circulation Model (OpenGGCM) (Raeder et al., 1998) (see Welling, 2019). The SWMF consists of several numerical modules, such as the ideal MHD solver BATS-R-US (Block Adaptive Tree Solar-wind Roe-type Upwind Scheme) (Powell et al., 1999; De Zeeuw et al., 2000; Gombosi et al., 2002), the Ridley Ionosphere Model (RIM) model (Ridley et al., 2002), and the inner magnetosphere Rice Convection Model (RCM) (Toffoletto et al., 2003).

The operational magnetospheric MHD models underestimate the magnitude of the perturbations across a wide frequency range, including the sub-hourly variations important for GICs (Welling, 2019). Pulkkinen et al. (2013) examined $\frac{d\mathbf{B}}{dt}$ on a 1-minute timescale and found an underestimation of magnitude between a factor of 2 and 10. Without the large $\frac{d\mathbf{B}}{dt}$ associated with high-frequency variations and resolution of peaks in the geomagnetic field, the magnitude of the derived surface geoelectric field ($\mathbf{E}$) is too low, resulting in an underestimation of GIC magnitudes.

However, a counter example is Raeder et al. (2001) who used an MHD model to simulate the Bastille day storm and compared their results to observations. Under a power spectral density (PSD) analysis they found that the model worked well for frequencies of 0-3 mHz and actually gave an overestimation at higher frequencies. These results are likely due to using a model configuration with a high grid resolution that would currently be prohibitive for operational forecasting, particularly if large ensembles of magnetospheric runs are required to estimate forecast uncertainty.

Figure 1 shows an example of SWMF power spectrum at a broad range of frequencies. The observed and modelled (using SWMF) horizontal magnetic field, the magnetic field component most relevant to GICs, is shown for the December 2006 CCMC test case (`https://ccmc.gsfc.nasa.gov/challenges/dBdt/`) at the Newport magnetometer site. The time series are shown in Figure 1 a) and the resulting power spectra in Figure 1 b). The coloured lines represent different model configurations. The power spectra shows that each configuration of the model underestimates the power spectral density, however the magnitude of underestimation is highly sensitive to model configuration with 12a_SWMF, the current operational configuration, performing best. These models are giving an output at a 1-minute resolution but the timeseries is smoother than that observed, meaning the amplitude of the higher frequency variations is reduced as shown by the power spectra. These simulation results have been provided by the Community Coordinated Modeling Center at Goddard Space Flight Center for the 2013 Space Weather Workshop and and an online interface is available for analysis of the model runs (`https://ccmc.gsfc.nasa.gov/challenges/dBdt/`).

A general underestimation is in agreement with Pulkkinen et al. (2013), who show in their Figures 3 and 4 that SWMF underestimated $\frac{d\mathbf{B}}{dt}$. Although we here only show that SWMF exhibits this underestimation, we note that this underestimation is a general feature of operational models predicting geomagnetic perturbations (Pulkkinen et al., 2010, 2011, 2013).

Recent work from Dimmock et al. (2021) tested different spatial resolution configurations of SWMF for the September 2017 event. They found that the high resolution made a significant improvement to the PSD and GIC forecasts. However, they noted that SWMF performs poorly in substorms and increasing the resolution has limited benefit in these periods. They concluded that a skilful GIC forecast can be done with SWMF but that computational power makes this operationally difficult. In contrast, Haiducek et al. (2017) compared the performance of SWMF on an event in 2005 using the reso-

lution of the operational model and a higher resolution. They used this configuration to estimate geomagnetic indices and cross-polar cap potential (CPCP). They found that results were not sensitive to resolution with the exception of predicting AL which may have been improved. The discrepancy is possibly because Haiducek et al. (2017) did not increase the resolution nearly as much as Dimmock et al. (2021). Mukhopadhyay et al. (2020) also used the configurations of Haiducek et al. (2017) finding that the high-resolution configuration performed generally better under the Heidke skill score for predicting $\frac{d\mathbf{B}}{dt}$.

Several further studies have shown that non-standard MHD model configurations can achieve excellent results for small scale phenomena in a statistical sense. Welling et al. (2021) modelled the magnetospheric response to a hypothetical "perfect" coronal mass ejection and successfully resolved high frequency phenomena. Realistic studies of ULF waves have been made by MHD models (Hartinger et al., 2014; Claudepierre et al., 2009) and small spatial and temporal features have been resolved by a new MHD model (Sorathia et al., 2020). These studies show that MHD models have the capability of properly capturing high frequency ground perturbations relevant to GICs, but the model configurations required are currently computationally prohibitive for operational real-time forecasting.

A viable operational alternative to increasing MHD model grid resolution is through the use of a method that statistically relates variability across temporal scales, namely a statistical downscaling approach. In addition to improving the geoelectic field reconstruction from a single magnetospheric model run, downscaling also has the potential to allow uncertainty quantification without the need for a magnetospheric model ensemble.

This paper addresses the characterisation of high-frequency variability in the magnetic field, $\mathbf{B}$, through statistical downscaling. Downscaling has been used in terrestrial weather forecasting to effectively increase the temporal and spatial resolution of global climate models (GCMs)(Maraun et al., 2010; Christensen & Christensen, 2003). For rainfall, this is done because rainfall typically occurs on subgrid scales so cannot be accurately captured with a GCM alone.

Maraun et al. (2010) classifies downscaling into three general categories: perfect prognosis approaches, model output statistics, and weather generators. Perfect prognosis approaches statistically determine relationships between low resolution predictors and the high resolution predictands. This works if the predictors are realistic, such as from a perfect (low resolution) forecast model, i.e. a perfect prognosis. Model output statistics builds a similar statistical relationship but with the aim of also correcting the bias of the forecast model. As such, model output statistics are model-specific. Finally, weather generators generate new high resolution time series that have the same statistical properties as observations, rather than just a probability of a sub grid event. Weather generators can be either perfect prognosis or model output statistics based.

As discussed by Morley (2020), statistical downscaling is relevant to space physics, in particular, to solar wind parameters used as inputs to magnetospheric models. Owens et al. (2014) considered temporal downscaling of solar wind parameters for this purpose. This was done because the magnetospheric models are sensitive to variability at a higher time resolution than is represented in numerical solar wind forecasts. Owens et al. (2014) used a random noise generator that gave high temporal noise with approximately correct statistical properties and added this noise onto the baseline of the solar wind parameters. They found that even relatively simple solar wind downscaling significantly increased the value of the subsequent magnetospheric forecast.

In this work we employ temporal downscaling to increase the variability of magnetic field time series on the ground. By developing a model-independent perfect prognosis scheme, we are assuming that future global MHD models will provide a perfect rep-

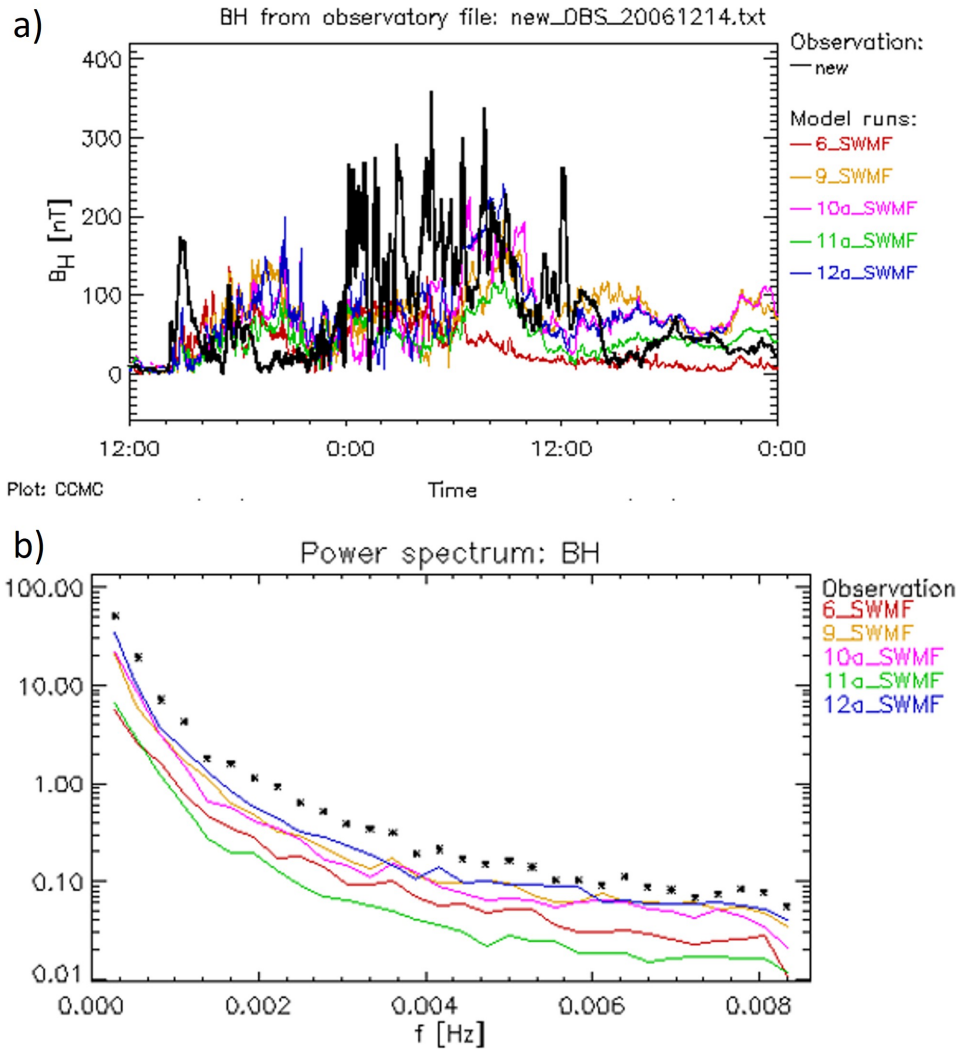


Figure 1: Geomagnetic field perturbations at Newport magnetometer station in December 2006. Several configurations of the SWMF (coloured lines) are compared with observations (black). a) shows the time series of the horizontal magnetic field. b) shows the associated power spectra for periods of 2 minutes and less revealing that SWMF underestimated the variability. These plots have been created and downloaded from the Community Coordinated Modelling Centre (CCMC) (`https://ccmc.gsfc.nasa.gov/challenges/dBdt/`.)

resentation of the low resolution magnetic field variations and/or model biases can be corrected by other means. However, the approach will be applicable to global MHD models that return a skilful and unbiased representation of the low resolution magnetic field. As the high-frequency variations are sampled from an ensemble of observations, an ensemble of geoelectric field estimates can also be reproduced from a single magnetospheric model run.

In the future we hope to apply our downscaling methodology directly to forecasts provided by global MHD models and potentially as a means for uncertainty estimation. However, it is important to develop and test the downscaling scheme in isolation, and not to convolve it with the performance of a specific magnetospheric model. Thus we

adopt the widely-used perfect prognostic approach and produce a perfect low-resolution forecast time series by taking 1-hour boxcar means of **B** observed by ground-based magnetometers. This 1-hour series is then linearly interpolated to 1-minute resolution. This represents the undownscaled time series. As will be shown in Section 4, this undownscaled series effectively removes all power in variations below 1 hour. Thus it is not a direct proxy for high-resolution magnetospheric model output. However, we start from this 1-hour linearly-interpolated undownscaled series for two reasons. Firstly, we expect magnetospheric models to perform better than this but it can be thought of as a 'worst-case scenario' for low-resolution magnetospheric models such as might be used for real-time forecasting in large ensembles. Secondly, if the downscaling manages to successfully relate the variability at 1-hour resolution to that at 1-minute resolution, it should be more than adequate for use with magnetospheric model output.

The downscaling scheme attempts to reintroduce high-frequency perturbations onto the linearly-interpolated 1-hour time series to produce a more realistic (in a statistical sense) **B** time series at the 1-minute resolution. By using observations as the undownscaled time series, rather than model output, we removing model error from the process of developing and testing our methodology. Additionally, this approach allows us to easily create a large database of low-resolution, undownscaled "forecasts" with which to test our model, without requiring decades of magnetospheric model output.

2 Data

The ground-based magnetometer measurements we use are provided by SuperMAG (Gjerloev, 2012) (`http://supermag.jhuapl.edu`), an international collaboration bringing together data from over 300 magnetometer stations. The SuperMAG ground-level magnetic field perturbation data has been homogenised in terms of coordinate system, processing technique and file structure.

A ground-based magnetometer measures the magnetic field from all sources in its vicinity. For studies on magnetic perturbations due to ionospheric and magnetospheric current systems, the magnetic baseline needs to be subtracted from the measurements to remove effects from other magnetic sources such as the Earth's intrinsic magnetic field. Gjerloev (2012) describes the SuperMAG data-processing technique for removing the base line, in which knowledge of typical timescales of variations of different magnetic fields is used. These amount to a yearly trend, mainly due to the secular variation in the Earth's main field, and a diurnal trend due to the Sq current system, the quiet day daily variation in ionospheric activity due to solar radiation. These are subtracted from the magnetometer measurements, leaving the prime source of variability as space-weather driven activity.

Of course, magnetometer measurements can occasionally have erroneous measurements. These usually take the form of a spike in activity for a single data point during an otherwise quiet period. These errors can sometimes get past the SuperMAG quality control and into the final datasets. The data used for this analysis is a SuperMAG dataset that has been cleaned for occasions where an error has exceeded the $99.97th$ percentile in terms of the change in the magnetic field with time as described in Rogers et al. (2020). The data may still have errors at lower levels of activity.

In this study we primarily use data from the Eskdalemuir (ESK) station located in southern Scotland with geographic coordinates of 55.314°N and 356.794°E. In principle, temporal downscaling techniques are applicable to all locations but we first test this one location where we have access to an established model for converting local magnetic field variations to geoelectric field variations, acknowledging the local ground conductivity conditions (Beggan et al., 2021). From the ESK station we have 1-minute **B** measurements for approximately 30 years, from 1983-2016.

3 Methodology

3.1 Analogue Ensemble

The Analogue Ensemble (AnEn) approach was originally used for terrestrial weather forecasting (e.g. van den Dool, 1989; Delle Monache et al., 2013), but has been far surpassed by physics-based models for that application. However, AnEn has more recently been employed in space and magnetospheric physics where the physical models are less accurate, largely from the limited availability of observations to completely characterise the necessary boundary conditions. In such a situation, empirical schemes can be valuable. Haines et al. (2021), Owens et al. (2017), Riley et al. (2017) and Barnard et al. (2011) have experimented with AnEn for forecasting the solar wind, geomagnetic activity and changes in space climate. In each case AnEn outperformed the benchmarks considered.

The AnEn methodology exploits an extensive historical dataset for forecasting purposes through analogy to past evolution of a given system. Specifically, an AnEn examines the present state of the predictors, looks in the historical dataset for analogous periods, then takes the predictand from the most analogous period. By selecting multiple analogous periods, an ensemble of predictands can be created, enabling a probabilistic forecast of future evolution.

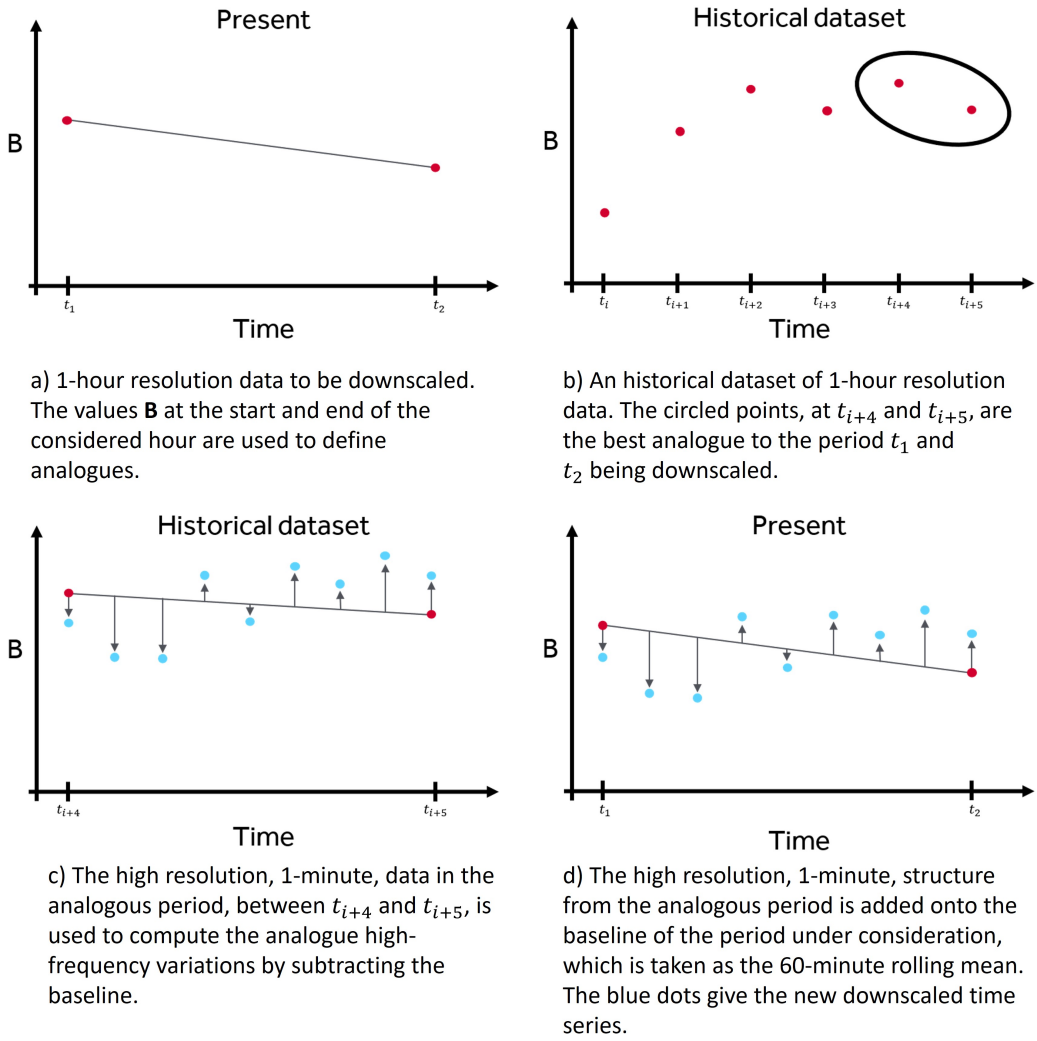
In this work, AnEn is used not for forecasting, but for temporal downscaling to relate variations on long and short timescales. To demonstrate that the downscaling framework works for ground-level **B**, we chose 1-hour and 1-minute for the long and short timescales somewhat arbitrarily, as described in the previous section. They are intended as examples rather than fixed parameters. At the high frequency, 1-minute makes sense as that is the typically available resolution of long-term ground-based **B** series and also the input resolution for many geoelectric field models. At the low frequency, the time scale of interest will depend on the specific model and the situation in which the model is being used. e.g., where real-time forecasting is required and/or ensembles of magnetospheric models are being used, it may be necessary to reduce the model resolution. As said, the low-resolution timescale of 1-hour is a tuneable parameter. If the downscaling is able to successfully relate 1-hour and 1-minute variations, it should perform even better at relating, e.g., 20-minute and 1-minute variations. Due to the perfect prognostic approach we can use the low-resolution time series as predictors. Specifically, the predictors used are the low-resolution values of the horizontal magnetic field at the start and the end of the considered hour. Analogous periods of these are found and used to predict a 1-minute resolution time series.

The AnEn algorithm is outlined in Figure 2 and described in the following points, in which the subscript H stands for 1-hour and M for 1-minute values:

1. Split the 1-minute SuperMAG data into two sets ($D1_M$, $D2_M$). $D1_M$ is the test dataset containing the short period to be downscaled. $D2_M$ is the independent training dataset comprised of the remaining data.
2. Compute low-resolution data using a 1-hour box-car means, to give $D1_H$ and $D2_H$.
3. Using $D1_H$, take the values at the start (t_1) and end (t_2) of the hour being considered, as shown in Figure 2 a).
4. Search $D2_H$ for the N most similar consecutive values, by mean squared error, to those at t_1 and t_2, as in Figure 2 b), where N is the chosen number of analogues.
5. Remove the baseline value from the associated $D2_M$ leaving only the higher frequency structure of the analogue interval, i.e. minute-scale variations with the baseline removed, as in Figure 2 c). The baseline is defined as the 60-minute rolling mean.
6. Add each $D2_M$ analogue onto $D1_H$ to produce an ensemble of downscaled values as in Figure 2 d).

7. Repeat this process for each hour in $D1_H$.

The data is then repeatedly split into different test and training sets so that the whole 34-year period can be downscaled using an independent training set. Note that this procedure uses data from after the 'forecast' time, so is not strictly a hindcast. However, this approach uses the volume of available historical data available to a forecast made today and thus quantifies the current expected performance of downscaling.



a) 1-hour resolution data to be downscaled. The values **B** at the start and end of the considered hour are used to define analogues.

b) An historical dataset of 1-hour resolution data. The circled points, at t_{i+4} and t_{i+5}, are the best analogue to the period t_1 and t_2 being downscaled.

c) The high resolution, 1-minute, data in the analogous period, between t_{i+4} and t_{i+5}, is used to compute the analogue high-frequency variations by subtracting the baseline.

d) The high resolution, 1-minute, structure from the analogous period is added onto the baseline of the period under consideration, which is taken as the 60-minute rolling mean. The blue dots give the new downscaled time series.

Figure 2: A schematic of the AnEn process. This process is repeated with the N best analogous periods to give an ensemble of downscaled time series.

3.2 Reference model

We use a reference model, as suggested by Liemohn et al. (2018), as a benchmark of comparison for the AnEn's performance. As this is a proof of concept study, we choose a reference model that represents a "do-nothing" approach to downscaling. For this we downscale the 1-hour time series of the magnetic field using a linear-interpolation, denoted as the linear-interpolation approach. Through this, we end up with 1-minute resolution time series without adding further high resolution structure.

As stated in Section 2, this 1-hour linear-interpolation series is not representative of ground-level **B** produced by typical state-of-the-art magnetospheric models, as can be seen from the power spectra in Figures 1 and 4. Instead, 1-hour can be seen more as a worst-case scenario – most magnetospheric models would be expected to reasonably reproduce the **B**-field fluctuation power at around 0.00028 Hz, even in real-time ensembles.

3.3 MT-transfer function

The goal of this work is not to recreate the high resolution magnetic field on a point-by-point basis, but to add in realistic high-frequency variability in a statistical sense. In particular, we are interested in the higher frequency structure insofar as it improves the subsequent estimate of the induced geoelectric field, which is the driver of GICs.

This can be tested with an "impacts" model. For this purpose we use a magnetotelluric-(MT-) transfer function (Simpson & Bahr, 2020; Beggan et al., 2021) produced for the ESK site by the British Geological Survey (BGS). The MT-transfer function converts a time series of the local magnetic field into a time series of local geoelectric field. The MT-transfer function first makes a Fourier transform of the magnetic field, then multiplies the result by an empirically determined matrix of coefficients which account for the local ground conductivity, and finally makes an inverse Fourier transform to compute the geoelectric field in the time domain. The matrix of coefficients is derived from simultaneous observations of the magnetic and geoelectric fields at ESK.

To quantify the performance of the downscaling scheme, we focus on the magnitude of the estimated **E**-field. Each **B**-field ensemble member was individually transformed with the MT-transfer function to result in an associated **E**-field ensemble member. A 'good' outcome would be that the $|\mathbf{E}|$ from the downscaled series is closer to the $|\mathbf{E}|$ obtained from using the observed series, than the linear-interpolation approach. An ideal outcome would be that the observed $|\mathbf{E}|$ output falls within the spread of the ensemble of $|\mathbf{E}|$ outputs obtained with the ensemble of downscaled series.

4 Evaluation

The AnEn downscaling approach has been applied to the entire 34-year period (1983-2014) of observations using an ensemble of 100 members built hour by hour as described above. Figure 3 shows an example spanning six-hours of heightened activity, with the x-component (east-west) in 3 a) and the y-component (north-south) in 3 b). This period was a geomagnetic storm with a minimum Dst of -172 nT. The observed time series is shown in red, the linear-interpolation series is shown in blue, and the median of the AnEn series is shown in black, with colour bands showing the $0th-100th$, $10th-90th$ and $25th-75th$ percentiles. The linear-interpolation approach is shown as a benchmark for the AnEn series to be compared against.

For the interval shown in Figure 3, the $10th-90th$ percentile band captures some of the variability seen in the observations, however, it seriously underestimates the variability on several occasions. Notably, towards the middle of the period, when the event is at the peak, the ensemble spread captures less of the variability. This suggests that the AnEn will struggle with the larger events such as this. By the definition of confidence, we would expect the observation to sit within the $0th-100th$ percentile band 100%of the time, in the $10th-90th$ percentile band 80% of the time and in the $25th-75th$ percentile band 50% of the time. In actuality here, the percentage of observations in the $0th-100th$, $10th-90th$ and $25th-75th$ percentile bands for B_x are 83.4%, 40.3% and 20.3%, respectively. For B_y this is 98.9, 51.8% and 21.3%, respectively for this illustrative period.

Figure 4 shows the power spectra of the magnetic field from observations and AnEn. Shown is the median and percentile bands of the PSD's achieved by all the ensemble members computed with Welch's method using the Hanning window without overlap. The AnEn ensemble follows the observations closely with a general trend to slightly underestimate the power at lower frequencies (0-0.003) and slightly overestimate the power for higher frequnecies (0.007 and above). The 10-90% range of the AnEn is very narrow at approximately 0.5 at the most, reflecting a consistent performance across the whole ensemble. The linear-interpolation approach is shown in blue but has been cut off because, as expected, the power spectral density is very low and hence makes scaling the y-axis difficult. It is clear that AnEn provides a power spectrum much more similar to that of the observations than the linear-interpolation approach achieves.

To measure the effectiveness of adding higher frequency structure we use the **B** time series magnetic fields from the observations, AnEn and the linear-interpolation approach to drive the MT-transfer function as described in Section 3.3. The output of the MT-transfer model is shown in Figure 5 for the same six-hour period shown in Figure 3. We see that the AnEn captures some of the geoelectric field variability within its spread but the observations lie outside the range of the analogue ensemble on many occasions. The percentage of observations in the $0th - 100th$, $10th - 90th$ and $25th - 75th$ percentile bands for E_x are 97.4%, 59.5% and 31.3%, respectively. For E_y this is 97.4%, 51.8% and 27.4%, respectively for this illustrative period.

Figure 5 reveals that, as expected, the linear-interpolation series yields very low geoelectric fields, without any significant variation. With a large ensemble size, the AnEn median will tend toward a smooth line despite variations in individual ensemble members. Therefore, the usefulness of AnEn is not in its median but rather in the spread of its ensemble members for showing possible realisations of the timeseries. Because of this it is not useful to directly compare AnEn median to the linear-interpolation approach values. However we do see that the spread on the analogue ensemble is of a more similar magnitude to that in observations than the linear-interpolation approach time series. In addition, AnEn provides an idea of the uncertainty in a forecast which is useful for making decisions.

While this example period is illustrative, it is necessary to evaluate AnEn as a downscaling model over the full 34-year period using a set of metrics. In the following evaluation we have taken care to chose metrics that are robust to timing errors, as we make the assumption that the spectral properties of fluctuations and the magnitude of the peaks are generally more important than the phasing for GIC impacts. This is also relevant since operations require a lead time of possible occurrence and an estimate of the severity of that occurrence as they cannot implement system wide mitigation in real-time. When comparing data on a point-by-point basis, timing errors, in which a defined event is correctly predicted to occur but at slightly the wrong time, will incur a double penalty by many common metrics (e.g. see Figure 8 of Owens, 2018). For example, accuracy, which gives a fraction of correct predictions across the whole dataset, will count the forecast as wrong when it predicts an event that doesn't occur at the exact time step and wrong when the forecast does not predict an event that is observed, even if the time step is off by just one step.

The sensitive values of GIC magnitude and timescales are dependent on the set up of individual transformers and the power grid configuration. For example, the size of geoelectric field that will cause a significant GIC is dependent on the ground conductivity in the region around the transformer. We use horizontal geoelectric field as a practical solution to provide a general evaluation of the method (Beamish et al., 2002), however transformers are sensitive to the individual E_x and E_y parameters, depending on grid configuration (Orr et al., 2021).

4.1 Threshold-exceedance prediction

In this subsection we evaluate each individual ensemble member within AnEn for its ability to give a binary prediction of an event at individual time steps. We examine three levels of activity for event classification using the magnitude of total horizontal geoelectric field, denoted $|\mathbf{E}|$, from the MT-transfer function. The magnitude of the total horizontal geoelectric field is shown for an illustrative period in Figure 6. The chosen thresholds for evaluation are the $99th$, $99.9th$ and $99.99th$ percentiles of the magnitude of the total horizontal geoelectric field from the MT-transfer function driven by observed magnetic field time series over the period 1983 to 2016. These are 22.3, 58.8 and 171.9 mV/km respectively and shown in Figure 6 by the horizontal dashed lines. For context, the peak geoelectric field magnitude for the March 1989 storm at ESK was 411.4 mV/km as computed using the MT-transfer function. It is worth noting that the system collapse experience during this geomagnetic storm occurred before the peak due to the rapid onset of a substorm (Boteler, 2019).

In order to allow for timing errors at the minute scale, we evaluate AnEn using the fraction skill score (FSS) (Roberts & Lean, 2008; Owens, 2018). The FSS is most commonly used to measure the fractional occurrence of events in a given spatial window. Here, we use FSS with a 60-minute temporal window and count the fraction of predicted time points which are classified as events, and the fraction of observed time points which are events, within the same time window. This is repeated for each ensemble member for time windows covering the whole dataset and the mean squared error (MSE) between the observed and predicted fraction time series is computed. This is repeated for a reference forecast, in this case the linear-interpolation series, and the FSS is taken as $1-(MSE_{forecast}/MSE_{reference})$. A perfect forecast would achieve $FSS = 1$, a forecast with no skill compared to the reference would achieve $FSS = 0$ and a forecast performing worse than the reference will achieve a negative score. FSS is most useful to end users who need to know if an event will occur within a given time window without the need for exact (in this case, to the minute) knowledge of when it will occur.

Figure 7 shows the FSS achieved for each of the 100 ensemble members across the entire dataset for each of the three event thresholds. Ensemble ID is ordered from best to worst analogues considered, where best means the 1-hour values in the analogous periods are most similar to present conditions by RMSE. For the $99th$ percentile threshold (panel a) we see that each ensemble member has a positive FSS, with an average value across the whole ensemble of 0.095, showing it outperforms the reference method. When considering events over the $99.9th$ percentile, Figure 7 b) again shows all ensemble members having a positive FSS with an average across the ensemble of 0.17. We also see a clear trend in which ensemble members based upon better analogues produce better FSS scores. The increased visibility of the trend for the $99.9th$ percentile compared to the $99th$ percentile suggests that at higher thresholds we are inherently considering rarer events, which reduces the number of good analogues available.

For events over the $99.99th$ percentile (panel c) the FSS is mainly positive for the first 50 ensemble members and approximately zero for the second 50. The mean FSS for the whole ensemble is 0.067. There is a very stark decrease in the skill of the ensemble members as the ensemble ID increases suggesting that for such a high threshold there are only around 30 to 50 good analogues for AnEn to work with. This finding can help inform a decision on an appropriate ensemble size for deployment. It also suggests that it would be appropriate to weight ensemble members if they are to be combined in any way.

4.2 1-hour mean value prediction

The impact of GICs on transformers can be dependent on time-integrated effects, meaning that problems occur when GICs exceed a certain threshold for a certain dura-

tion. With this in mind, we now evaluate the model using events classified using thresholds of the 1-hour mean value of $|\mathbf{E}|$ previously used. The hourly mean of the magnitude of geoelectric field is shown for an illustrative period is shown in Figure 8. We again consider thresholds at the $99th$, $99.9th$ and $99.99th$ percentiles of the 1-hour means of the horizontal geoelectric field magnitude from the observed time series. These values are 17.9, 47.0 and 139.0 mV/km respectively. These are shown on Figure 8 by the horizontal dashed lines. For context, the peak hourly mean observed at ESK during the March 1989 storm was 77.1 mV/km, suggesting that although the peaks of this storm were large, they were short lived. These metrics are useful as impacts of a heightened geoelectric field are often caused by sustained heightened values on approximately the tens of minutes to 1-hour time scale (Pulkkinen et al., 2017). The metrics in this section are useful to end users who need to know when periods of heightened activity will occur and users who are impacted by time-integrated effects.

4.2.1 Deterministic prediction

The first metric chosen is the Heidke skill score (HSS) (Jolliffe & Stephenson, 2003). HSS measures the accuracy of a model, taking into account the number of correct random forecasts. This allows for proper measurement of skill in a situation where an event is rare. In fact, the rarer the event considered, the less HSS takes into account correct predictions of "no event", which becomes the overwhelming majority class. HSS uses the four categories on a standard contingency table: the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) events. HSS is then given by:

$$HSS = \frac{TP + TN - crf}{TP + TN + FP + FN - crf}, \tag{1}$$

where crf, the number of correct random forecasts, is

$$crf = \frac{(TP + FP)(TP + FN) + (FP + TN)(FN + TN)}{n}, \tag{2}$$

where n is the total number of predictions.

HSS of AnEn is shown in Figure 9 for the three event thresholds considered. HSS is has been computed for each ensemble member shown by the yellow bars and HSS for the linear-interpolation approach is shown by the black dashed horizontal line. AnEn clearly outperforms the linear-interpolation approach and it generally achieves a good positive score with the exception of some of the ensemble members based on weaker analogues for the $99.99th$ percentile threshold. This again suggests that the available data set is too small for 100 analogues of more extreme events.

4.2.2 Probabilistic prediction

Next we evaluate AnEn in its ability to give a probabilistic prediction of an event by counting how many of the ensemble members predict an event and normalising by the size of the ensemble. This is evaluated using the Cost/Loss analysis (Murphy, 1977; Richardson, 2000; Owens et al., 2017), which allows different end users of a forecast to assess its value for their particular use case. The idea is that taking mitigating action due to a forecast incurs a Cost, C, of fixed value, and experiencing an event without taking mitigating action incurs a Loss, L, of fixed value. The Cost/Loss analysis sums these Costs and Losses for acting on a particular forecast across a long time series and compares the sum to that of a perfect forecast and a climatological forecast method (which, at all times, predicts the probability of an event as the fraction of time in which that event is experienced across the whole dataset). The result is the potential economic value (PEV) which is 1 for a perfect forecast, 0 for a forecast of equal ability to climatology, and negative for a forecast with worse ability than the climatology. PEV is given as a function

Threshold (percentile)	BSS (100 members)	BSS (20 members)
99th	0.30	0.32
99.9th	0.32	0.38
99.99th	0.15	0.31

Table 1: Brier skill score (BSS) for AnEn using the linear-interpolation approach as the reference. Three event thresholds are considered.

of the Cost/Loss ratio, C/L, which is between 0 and 1 for all end users that may find a forecast valuable. In a probabilistic Cost/Loss analysis that we employ here, mitigating action is taken if the probability given by AnEn exceeds the Cost/Loss ratio of the end user. For more details see Murphy (1977); Richardson (2000).

Figure 10 shows the PEV for the Cost/Loss domain $(0, 1)$ for the probabilistic downscaling from the AnEn and the linear-interpolation approach. We see that for all three event thresholds AnEn outperforms the reference method. We also see that the PEV is highest for the lower end of the Cost/Loss domain which means it will most benefit end users who better tolerate false alarms (false positives) rather than missed events (false negatives). This is because at the low end of the C/L domain the cost of taking mitigating action is very low compared to the loss incurred due to not taking action and an event happening. Therefore, these users would generally prefer to take mitigating action on a false alarm than not take action on a real event.

Finally we look at how AnEn performs under the Brier skill score (BSS) (Jolliffe & Stephenson, 2003). Like Cost/Loss analysis, BSS can compare probabilistic forecasts with deterministic ones, allowing direct comparison of the probabilistic AnEn and the deterministic undownscaled series. BSS is useful to end users who wish to use the probabilistic information of AnEn. To compute BSS, the standard Brier score (BS) must first be computed. The BS is the normalised sum of the square error between the probabilistic forecast and the observations over the whole time series, where the observations takes a binary value of 0 or 1 depending on whether an event occurs. Events are again taken to be hours exceeding the $99th$, $99.9th$ and $99.99th$ percentiles of observed $|\mathbf{E}|$. BS is computed for both AnEn and the reference model then combined into BSS by

$$BSS = 1 - \frac{BS_{forecast}}{BS_{reference}}. \tag{3}$$

Similarly to the Cost/Loss and FSS, a perfectly skilful forecast receives BSS=1, a forecast with no skill relative to the reference receives BSS=0, and a negative score signals the forecast method performs worse than the reference.

BSS is shown for AnEn for the three event thresholds in Table 1. It seems that the 100-member AnEn has skill over the linear-interpolation approach for all considered thresholds but drops in skill for the $99.99th$ percentile events. It is likely that this is the result of the limited span of the dataset and hence number of analogous extreme events. A reduced ensemble size or ensemble-member weighting would likely yield a better BSS, particularly for the $99.99th$ percentile events. This is shown in the third column of the table which gives BSS for a 20 member ensemble. We see that the BSS of the $99.99th$ percentile events increases more in line with the lower thresholds.

5 Discussion & Conclusions

Statistical downscaling of magnetic field data for the purposes of GIC forecasting has been demonstrated in the form of a perfect prognostic approach. We employed the analogue ensemble (AnEn) methodology, finding that with its spread and higher frequency contributions, a more accurate E-field mapping is obtained than when compared to an E-field derived from undownscaled **B**-field data.

To obtain a "low-resolution" dataset, ground-level magnetic field perturbation data was smoothed from high frequency (1-minute) to low frequency (1-hour) resolution. High frequency structure was then reintroduced into the low-resolution (1-hour) series using the AnEn approach. Both the low frequency and the downscaled time series were then used in a magnetotelluric-transfer function to compute the corresponding horizontal geoelectric fields.

We presented the power spectrum of the observations, AnEn showing that AnEn closely resembles the spectral properties of the observations and far outperforms the linear-interpolation approach. Although AnEn has not been applied to the output of a global MHD model, it can be seen that it has the potential to improve the spectral properties of a forecast that has an underestimation of spectral power at the high frequencies.

The method was validated using a range of methods to test different aspects of the downscaling scheme. Specifically, we used the fraction skill score (FSS), Heidke skill score (HSS), Cost/Loss analysis and Brier skill score (BSS). FSS was used to evaluate AnEn on the occurrence rate of 1-minute events within 1-hour windows. The events were defined using three thresholds, namely, $99th$, $99.9th$ and $99.99th$ percentile of the entire dataset (1983 to 2016). AnEn had a positive FSS for all ensemble members for the $99th$, $99.9th$ percentile thresholds showing that AnEn outperformed the undownscaled approach. For the $99.99th$ percentile threshold, some of the weaker analogues achieved a negative FSS suggesting that the ensemble size of 100 was too large for the current dataset to allow good analogues of the most extreme events to be found. Nevertheless, the overall FSS was still positive.

Since impacts of GICs tend to require an elevated geoelectric field over a sustained period, we also evaluated AnEn for its ability to predict the hourly-mean value of geoelectric field. This was achieved by defining events as the 1-hour mean value exceeding the thresholds of $99th$, $99.9th$ and $99.99th$ percentile of the hourly-means of the entire dataset. With this event definition, HSS revealed that AnEn outperformed the undownscaled series for all ensemble members in the three event thresholds, except for a small number in the $99.99th$ percentile events.

This work has evaluated AnEn with an ensemble size of 100. The ensemble size should be chosen large enough that a wide range of possible outcomes can be included, but small enough to ensure analogues are of a good quality and are in fact analogous. The fraction skill score and Heidke skill score revealed that better quality analogues downscaled more skilfully. The number of good quality analogues available depends both on the size of the historical dataset and on the rarity of event considered. This was particularly evident when considering events above the $99.99th$ percentile suggesting 100 members is too many to ensure all analogues are of a good quality. A more appropriate ensemble size for this threshold would be approximately 20 as shown by the BSS analysis. Future implementations of this method should use these results to inform an appropriate ensemble size for the size of event of interest.

In this work the probabilistic prediction given by AnEn was made by simple ensemble member voting. The impact of analogue quality could be mitigated if, when converting to a probabilistic prediction from an ensemble of predictions, the voting power of each member is dependent on the quality of the analogue, as measure by the inverse of the RMSE between analogue and period under consideration and normalising. This

	Eskdalemuir			Lerwick			Hartland		
	99	99.9	99.99	99	99.9	99.99	99	99.9	99.99
Mean FSS	0.10	0.17	0.07	0.41	0.28	0.07	0.11	-0.06	-0.04
Mean HSS	0.31	0.32	0.12	0.54	0.34	0.13	0.27	0.17	0.03
BSS	0.30	0.32	0.15	0.51	0.35	0.14	0.26	0.15	0.04

Table 2: The mean FSS, mean HSS and BSS for the three thresholds at Eskdalemuir, Lerwick and Hartland.

would mean that members expected to have the most insight into the situation have greater sway in the overall prediction.

We implemented a probabilistic Cost/Loss analysis revealing that AnEn has a higher potential economic value than the undownscaled approach and that the value of the forecast was greater for end users who can tolerate false alarms at the lower end of the Cost/Loss domain. Like the previous metrics, AnEn performed better in the $99th$ and $99.9th$ percentile events.

A shortcoming of AnEn is that there is expected to be a lack of good analogues for the most extreme events. To address this AnEn could be improved by expanding the predictors used to include such things as geomagnetic indices and estimates of current systems. This could allow AnEn to be more aware of the drivers of geomagnetic activity and thus allow the use of fewer-but-better-quality analogues in a reduced size ensemble. Although this is a shortcoming, it is important to remember moderate space weather events are problematic as well as the rarer, more extreme events (e.g. Schrijver et al., 2014; Schrijver, 2015). A further way to increase ensemble member quality would be to create the training dataset, $D2_M$, using a rolling-mean rather than box-car as this would create a more potential analogous periods and hence increase analogue quality overall.

We used a perfect prognostic approach to downscaling which assumes the low time resolution forecast given is a perfect forecast. This allowed us to use historical observations as if they were forecast model outputs. However, this approach is limited because the models are not perfect. It is expected that biases in the forecast model would not be corrected but carried through by the downscaling methodology.

This paper has focused on the results for the Eskdalemuir station, however, an equivalent analysis has been conducted for the Lerwick and Hartland magnetometer stations in the UK. The AnEn downscaling methodology applied to these stations generally perform similarly to ESK, supporting the claim that this methodology could be applied more broadly. The achieved mean FSS, mean HSS and BSS for events above the three thresholds are shown in Table 2 for Lerwick and Hartland. The results for ESK are also shown for reference. AnEn is shown to perform to a slightly better standard at Lerwick, particularly for the $99th$ percentile threshold, and slightly worse at Hartland, particularly for the higher thresholds.

In this work, AnEn has been used both to generate a downscaled time series and to estimate the uncertainty of it by using many ensemble members. It would be quite possible to remove the downscaling element and just use the algorithm to provide probabilistic information for a forecast that already has the correct spectral properties.

This work has given proof of concept that downscaling can be implemented to improve a forecast that lacks realistic high-frequency structure. From here, research should be conducted to create downscaling schemes that are optimised to perform better than

AnEn when the downscaled data is used to drive an "impacts" model. The optimisation could include finding different model configurations for specific space weather drivers. This would take knowledge of the solar wind driving the magnetosphere, such as CMEs or CIRs, and restrict AnEn to choosing analogues from historical periods driven by the same solar wind context. Once downscaling methods have been further investigated, the front runners will need to be manipulated to form a "bolt-on" piece for a global MHD model. We finally note that the methods developed here do not attempt to correct for any biases in the magnetospheric models. Thus it remains to be seen whether the improvements demonstrated here translate directly to a forecasting situation, or where further bias-correction of magnetospheric models is also required.

Acknowledgments

The authors thank the National Environmental Research Council (NERC) for funding this work under grants NE/L002566/1 and NE/P016928/1.

For the ground magnetometer data we gratefully acknowledge: INTERMAGNET, Alan Thomson; CARISMA, PI Ian Mann; CANMOS, Geomagnetism Unit of the Geological Survey of Canada; The S-RAMP Database, PI K. Yumoto and Dr. K. Shiokawa; The SPIDR database; AARI, PI Oleg Troshichev; The MACCS program, PI M. Engebretson; GIMA; MEASURE, UCLA IGPP and Florida Institute of Technology; SAMBA, PI Eftyhia Zesta; 210 Chain, PI K. Yumoto; SAMNET, PI Farideh Honary; IMAGE, PI Liisa Juusola; Finnish Meteorological Institute, PI Liisa Juusola; Sodankylä Geophysical Observatory, PI Tero Raita; UiT the Arctic University of Norway, Tromsø Geophysical Observatory, PI Magnar G. Johnsen; GFZ German Research Centre For Geosciences, PI Jürgen Matzka; Institute of Geophysics, Polish Academy of Sciences, PI Anne Neska and Jan Reda; Polar Geophysical Institute, PI Alexander Yahnin and Yarolav Sakharov; Geological Survey of Sweden, PI Gerhard Schwarz; Swedish Institute of Space Physics, PI Masatoshi Yamauchi; AUTUMN, PI Martin Connors; DTU Space, Thom Edwards and PI Anna Willer; South Pole and McMurdo Magnetometer, PI's Louis J. Lanzarotti and Alan T. Weatherwax; ICESTAR; RAPIDMAG; British Artarctic Survey; McMac, PI Dr. Peter Chi; BGS, PI Dr. Susan Macmillan; Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation (IZMIRAN); MFGI, PI B. Heilig; Institute of Geophysics, Polish Academy of Sciences, PI Anne Neska and Jan Reda; University of L'Aquila, PI M. Vellante; BCMT, V. Lesur and A. Chambodut; Data obtained in cooperation with Geoscience Australia, PI Andrew Lewis; AALPIP, co-PIs Bob Clauer and Michael Hartinger; MagStar, PI Jennifer Gannon; SuperMAG, PI Jesper W. Gjerloev; Data obtained in cooperation with the Australian Bureau of Meteorology, PI Richard Marshall. The SuperMAG data is available at `https://supermag.jhuapl.edu/`.

Simulation results have been provided by the Community Coordinated Modeling Center at Goddard Space Flight Center through their public Runs on Request system (http://ccmc.gsfc.nasa.gov). This work was carried out using the SWMF and BATS-R-US tools developed at the University of Michigan's Center for Space Environment Modeling (CSEM). The modeling tools described in this publication are available online through the University of Michigan for download and are available for use at the Community Coordinated Modeling Center (CCMC). These simulation results can be found from `https://ccmc.gsfc.nasa.gov/challenges/dBdt/`.

References

Barnard, L., Lockwood, M., Hapgood, M. A., Owens, M. J., Davis, C. J., & Steinhilber, F. (2011). Predicting space climate change. *Geophysical Research Letters*, *38*(16), 7–12. doi: 10.1029/2011GL048489

Beamish, D., Clark, T. D., Clarke, E., & Thomson, A. W. (2002). Geomagnetically

induced currents in the UK: Geomagnetic variations and surface electric fields. *Journal of Atmospheric and Solar-Terrestrial Physics*, *64*(16), 1779–1792. doi: 10.1016/S1364-6826(02)00127-X

Beggan, C. D., Richardson, G. S., Baillie, O., Hubert, J., & Thomson, A. W. P. (2021). Geoelectric field measurement, modelling and validation during geomagnetic storms in the UK. *Journal of Space Weather and Space Climate*.

Boteler, D. H. (2019). A 21st Century View of the March 1989 Magnetic Storm. *Space Weather*, *17*(10), 1427–1441. Retrieved from `https://doi.org/10.1029/2019SW002278` doi: 10.1029/2019SW002278

Cannon, P., Angling, M., Barclay, L., Curry, C., Dyer, C., Edwards, R., ... Underwood, C. (2013). Extreme space weather: impacts on engineered systems and infrastructures. *Royal Academy of Engineering*, 70. doi: 1-903496-96-9

Christensen, J. H., & Christensen, O. B. (2003). Severe summertime flooding in Europe. *Nature*, *421*(6925), 805–806. doi: 10.1038/421805a

Claudepierre, S. G., Wiltberger, M., Elkington, S. R., Lotko, W., & Hudson, M. K. (2009). Magnetospheric cavity modes driven by solar wind dynamic pressure fluctuations. *Geophysical Research Letters*, *36*(13), 1–5. doi: 10.1029/2009GL039045

De Zeeuw, D. L., Gombosi, T. I., Groth, C. P., Powell, K. G., & Stout, Q. F. (2000). An adaptive MHD method for global space weather simulations. *IEEE Transactions on Plasma Science*, *28*(6), 1956–1965. doi: 10.1109/27.902224

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., & Searight, K. (2013). Probabilistic Weather Prediction with an Analog Ensemble. *Monthly Weather Review*, *141*(10), 3498–3516. doi: 10.1175/mwr-d-12-00281.1

Dimmock, A. P., Welling, D. T., Rosenqvist, L., Forsyth, C., Freeman, M. P., Rae, I. J., ... Yordanova, E. (2021). Modeling the Geomagnetic Response to the September 2017 Space Weather Event Over Fennoscandia Using the Space Weather Modeling Framework: Studying the Impacts of Spatial Resolution. *Space Weather*, *19*(5), 1–25. doi: 10.1029/2020sw002683

Gjerloev, J. W. (2012). The SuperMAG data processing technique. *Journal of Geophysical Research: Space Physics*, *117*(9), 1–19. doi: 10.1029/2012JA017683

Gombosi, T. I., Tóth, G., De Zeeuw, D. L., Hansen, K. C., Kabin, K., & Powell, K. G. (2002). Semirelativistic magnetohydrodynamics and physics-based convergence acceleration. *Journal of Computational Physics*, *177*(1), 176–205. doi: 10.1006/jcph.2002.7009

Haiducek, J. D., Welling, D. T., Ganushkina, N. Y., Morley, S. K., & Ozturk, D. S. (2017). SWMF Global Magnetosphere Simulations of January 2005: Geomagnetic Indices and Cross-Polar Cap Potential. *Space Weather*, *15*(12), 1567–1587. doi: 10.1002/2017SW001695

Haines, C., Owens, M. J., Barnard, L., Lockwood, M., & Ruffenach, A. (2021). Forecasting Occurrence and Intensity of Geomagnetic Activity with Pattern-Matching Approaches. *Submitted to Space Weather*, 1–23. doi: 10.1029/2020SW002624

Hartinger, M. D., Welling, D., Viall, N. M., Moldwin, M. B., & Ridley, A. (2014). The effect of magnetopause motion on fast mode resonance. *Journal of Geophysical Research: Space Physics*, *119*(10), 8212–8227. doi: 10.1002/2014JA020401

Jolliffe, I., & Stephenson, D. (2003). *Forecast verification: a practioners guide in atmospheric science* (Vol. 4) (No. 1).

Koskinen, H. E., Baker, D. N., Balogh, A., Gombosi, T., Veronig, A., & von Steiger, R. (2017). Achievements and Challenges in the Science of Space Weather. *Space Science Reviews*, *212*(3-4), 1137–1157. doi: 10.1007/s11214-017-0390-4

Liemohn, M. W., McCollough, J. P., Jordanova, V. K., Ngwira, C. M., Morley, S. K., Cid, C., ... Vasile, R. (2018). Model Evaluation Guidelines for Geomagnetic Index Predictions. *Space Weather*, *16*(12), 2079–2102. doi: 10.1029/2018SW002067

Lyon, J. G., Fedder, J. A., & Mobarry, C. M. (2004). The Lyon-Fedder-Mobarry

(LFM) global MHD magnetospheric simulation code. *Journal of Atmospheric and Solar-Terrestrial Physics*, *66*(15-16 SPEC. ISS.), 1333–1350. doi: 10.1016/j.jastp.2004.03.020

Maraun, D., Brienen, S., Rust, H. W., Sauter, T., Themeßl, M., Venema, V. K. C., & Chun, K. P. (2010). Precipitation Downscaling under Climate Change. *October*(2009), 1–34. doi: 10.1029/2009RG000314

Merkin, V. G., Owens, M. J., Spence, H. E., Hughes, W. J., & Quinn, J. M. (2007). Predicting magnetospheric dynamics with a coupled Sun-to-Earth model: Challenges and first results. *Space Weather*, *5*(12), 1–13. doi: 10.1029/2007SW000335

Morley, S. K. (2020). Challenges and Opportunities in Magnetospheric Space Weather Prediction. *Space Weather*, *18*(3). doi: 10.1029/2018SW002108

Mukhopadhyay, A., Welling, D. T., Liemohn, M. W., Ridley, A. J., Chakraborty, S., & Anderson, B. J. (2020). Conductance Model for Extreme Events: Impact of Auroral Conductance on Space Weather Forecasts. *Space Weather*, *18*(11), 1–27. doi: 10.1029/2020SW002551

Murphy, A. H. (1977). The Value of Climatological, Categorical and Probabilistic Forecasts in the Cost-Loss Ratio Situation. *Monthly Weather Review*, *105*(7), 803–816. doi: 10.1175/1520-0493

Orr, L., Chapman, S. C., & Beggan, C. (2021). Wavelet and network analysis of magnetic field variation and geomagnetically induced currents during large storms. *Space Weather*, 1–29. doi: 10.1029/2021sw002772

Oughton, E. J., Skelton, A., Horne, R. B., Thomson, A. W., & Gaunt, C. T. (2017). Quantifying the daily economic impact of extreme space weather due to failure in electricity transmission infrastructure. *Space Weather*, *15*(1), 65–83. doi: 10.1002/2016SW001491

Owens, M. J. (2018). Time-Window Approaches to Space-Weather Forecast Metrics: A Solar Wind Case Study. *Space Weather*, *16*(11), 1847–1861. doi: 10.1029/2018SW002059

Owens, M. J., Horbury, T. S., Wicks, R. T., McGregor, S. L., Savani, N. P., & Xiong, M. (2014). Ensemble downscaling in coupled solar wind-magnetosphere modeling for space weather forecasting. *Space Weather*, *12*(6), 395–405. doi: 10.1002/2014SW001064

Owens, M. J., Riley, P., & Horbury, T. S. (2017). Probabilistic Solar Wind and Geomagnetic Forecasting Using an Analogue Ensemble or "Similar Day" Approach. *Solar Physics*, *292*(5), 1–16. doi: 10.1007/s11207-017-1090-7

Powell, K. G., Roe, P., Linde, T., Gombosi, T., & De Zeeuw, D. L. (1999). A solution-adaptive scheme for ideal magnetohydrodynamics. *Computational Physics*, *154*, 284–309. doi: 10.1006/jcph.1999.6299

Pulkkinen, A., Bernabeu, E., Thomson, A., Viljanen, A., Pirjola, R., Boteler, D., ... MacAlester, M. (2017). Geomagnetically induced currents: Science, engineering, and applications readiness. *Space Weather*, *15*(7), 828–856. doi: 10.1002/2016SW001501

Pulkkinen, A., Kuznetsova, M., Ridley, A., Raeder, J., Vapirev, A., Weimer, D., ... Chulaki, A. (2011). Geospace Environment Modeling 2008-2009 Challenge: Ground magnetic field perturbations. *Space Weather*, *9*(2), 1–13. doi: 10.1029/2010SW000600

Pulkkinen, A., Rastätter, L., Kuznetsova, M., Hesse, M., Ridley, A., Raeder, J., ... Chulaki, A. (2010). Systematic evaluation of ground and geostationary magnetic field predictions generated by global magnetohydrodynamic models. *Journal of Geophysical Research: Space Physics*, *115*(3), 1–12. doi: 10.1029/2009JA014537

Pulkkinen, A., Rastätter, L., Kuznetsova, M., Singer, H., Balch, C., Weimer, D., ... Weigel, R. (2013). Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations.

Space Weather, *11*(6), 369–385. doi: 10.1002/swe.20056

Raeder, J., Berchem, J., & Ashour-Abdalla, M. (1998). The Geospace Environment Modeling Grand Challenge: Results from a Global Geospace Circulation Model. *Journal of Geophysical Research: Space Physics*, *103*(A7), 14787–14797. doi: 10.1029/98ja00014

Raeder, J., Wang, Y. L., Fuller-Rowell, T. J., & Singer, H. J. (2001). Global simulation of magnetospheric space weather effects of the Bastille day storm. *Solar Physics*, *204*(1-2), 325–338. doi: 10.1023/A:1014228230714

Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensamble prediction system. (October 1998), 649–667. doi: 10.1002/qj.49712656313

Ridley, A. J., Hansen, K. C., Tóth, G., De Zeeuw, D. L., Gombosi, T. I., & Powell, K. G. (2002). University of Michigan MHD results of the geospace global circulation model metrics challenge. *Journal of Geophysical Research: Space Physics*, *107*(A10), 1–19. doi: 10.1029/2001JA000253

Riley, P., Ben-Nun, M., Linker, J. A., Owens, M. J., & Horbury, T. S. (2017). Forecasting the properties of the solar wind using simple pattern recognition. *Space Weather*, *15*(3), 526–540. doi: 10.1002/2016SW001589

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 1–19. doi: 10.1175/2007MWR2123.1

Rogers, N. C., Wild, J. A., Eastoe, E. F., Gjerloev, J. W., & Thomson, A. W. (2020). A global climatological model of extreme geomagnetic field fluctuations. *Journal of Space Weather and Space Climate*, *10*, 1–19. doi: 10.1051/swsc/2020008

Schrijver, C. J. (2015). Socio-Economic Hazards and Impacts of Space Weather: The Important Range between Mild and Extreme. *Space Weather*, *13*(9), 524–528. doi: 10.1002/2015SW001252

Schrijver, C. J., Dobbins, R., Murtagh, W., & Petrinec, S. M. (2014). Assessing the impact of space weather on the electric power grid based on insurance claims for industrial electrical equipment. *Space Weather*, *12*(7), 487–498. doi: 10.1002/2014SW001066

Simpson, F., & Bahr, K. (2020). Nowcasting and validating Earth's electric-field response to extreme space-weather events using magnetotelluric data: application to the September 2017 geomagnetic storm and comparison to observed and modelled fields in Scotland. *Space Weather*(September), 0–1. doi: 10.1029/2019sw002432

Sorathia, K. A., Merkin, V. G., Panov, E. V., Zhang, B., Lyon, J. G., Garretson, J., ... Wiltberger, M. (2020). Ballooning-Interchange Instability in the Near-Earth Plasma Sheet and Auroral Beads: Global Magnetospheric Modeling at the Limit of the MHD Approximation. *Geophysical Research Letters*, *47*(14). doi: 10.1029/2020GL088227

Toffoletto, F., Sazykin, S., Spiro, R., & Wolf, R. (2003). Inner Magnetospheric Modeling with the Rice Convection Model. In *Advances in space environment research.* Springer. doi: 10.1007/978-94-007-1069-6_19

Tóth, G., Sokolov, I. V., Gombosi, T. I., Chesney, D. R., Clauer, C. R., De Zeeuw, D. L., ... Kóta, J. (2005). Space weather modeling framework: A new tool for the space science community. *Journal of Geophysical Research: Space Physics*, *110*(A12), 1–21. doi: 10.1029/2005JA011126

Tóth, G., van der Holst, B., Sokolov, I. V., De Zeeuw, D. L., Gombosi, T. I., Fang, F., ... Opher, M. (2012). Adaptive numerical algorithms in space weather modeling. *Journal of Computational Physics*, *231*(3), 870–903. doi: 10.1016/j.jcp.2011.02.006

van den Dool, H. M. (1989). A New Look at Weather Forecasting through Analogues. *Monthly Weather Review*, *117*(10), 2230–2247. doi: 10.1175/1520-0493(1989)117⟨2230:ANLAWF⟩2.0.CO;2

Weimer, D. R. (2013). An empirical model of ground-level geomagnetic perturbations. *Space Weather*, *11*(3), 107–120. doi: 10.1002/swe.20030

Weimer, D. R. (2019). Empirical modeling of the geomagnetic field for GIC prediction. In *Geomagnetically induced currents from the sun to the power grid* (pp. 67–78). doi: 10.1002/9781119434412.ch4

Welling, D. (2019). Magnetohydrodynamic models of B and their use in GIC estimates. In *Geomagnetically induced currents from the sun to the power grid* (pp. 43–65). AGU publications. doi: 10.1002/9781119434412.ch3

Welling, D., Love, J., Rigler, J., Oliveira, D., Komar, C., & Morley, S. (2021). Numerical Simulations of the Geospace Response to the Arrival of an Idealized Perfect Interplanetary Coronal Mass Ejection. *Space Weather*, *19*(2), 1–15. doi: 10.1029/2020sw002489

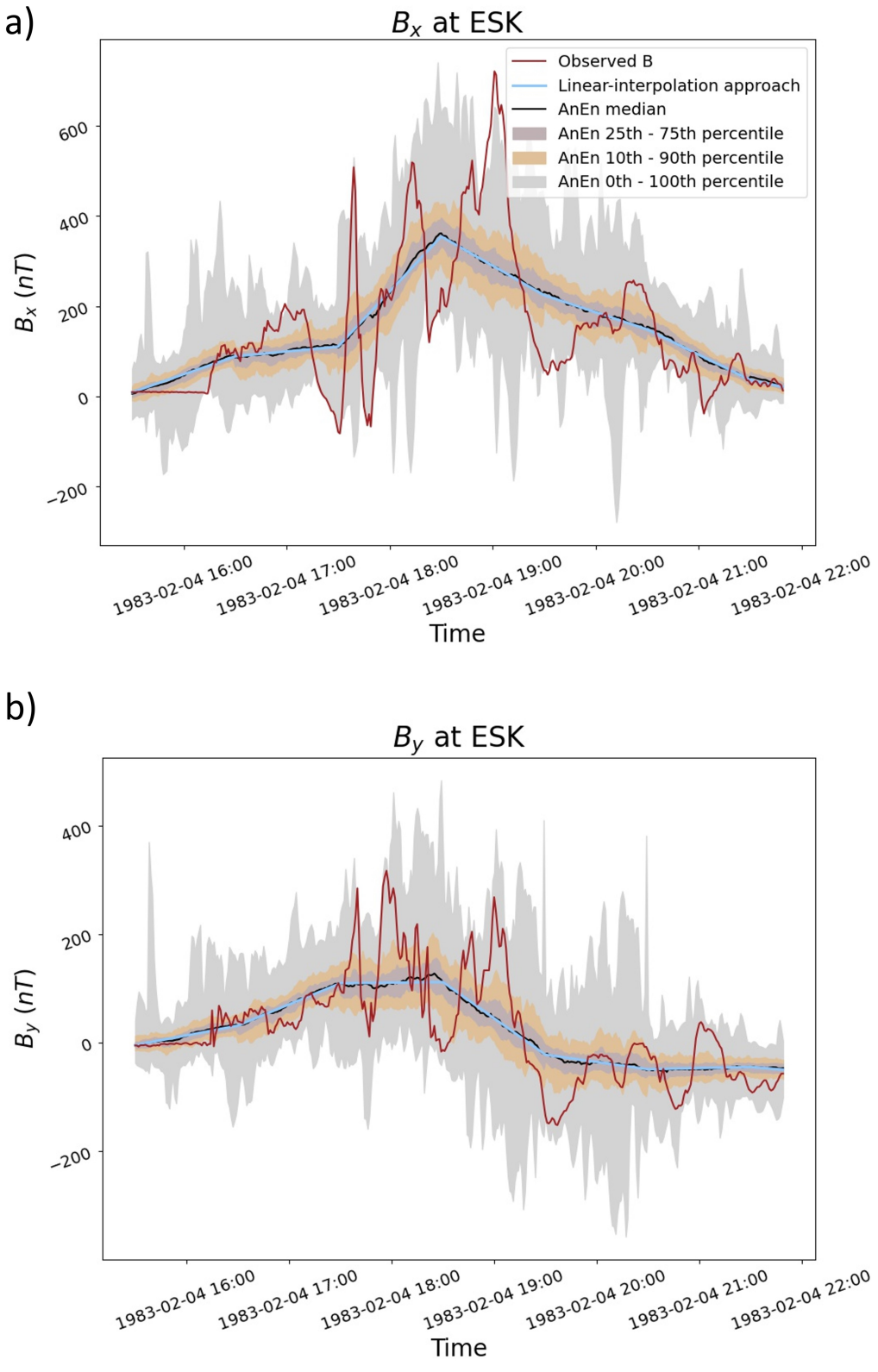


Figure 3: A six-hour time series from 1983-02-04 of the magnetic field at ESK in the x (east-west) and y (north-south) directions in the geographic coordinate system. The red line shows the observed 1-minute time series, the colour bands show the spread of the AnEn series (the 10th-90th and 25th-75th percentiles) with the median in black, and the blue line shows the linear-interpolation approach, taken to be the undownscaled magnetic field, as a reference.

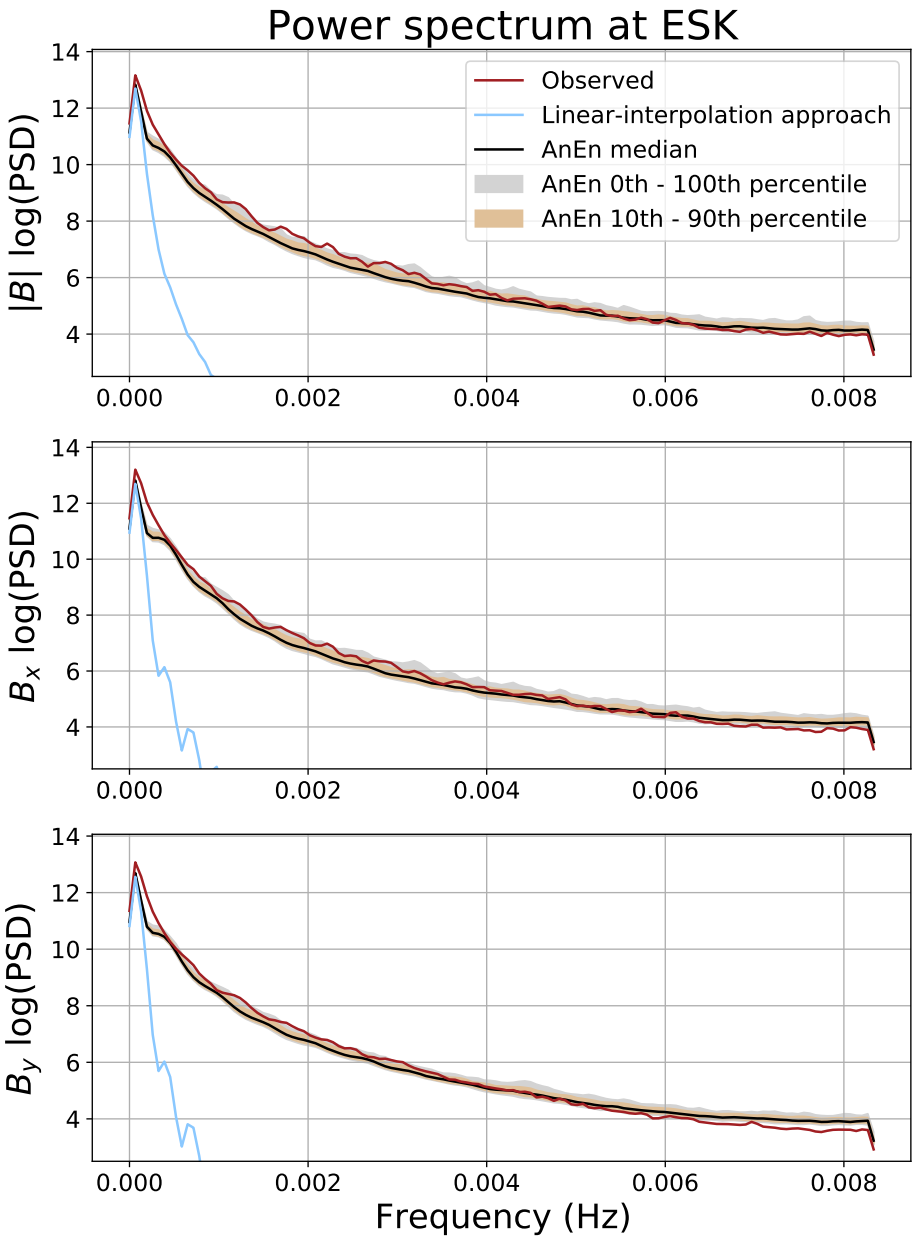


Figure 4: The power spectrum of the magnitude, B_x and B_y components of the magnetic field from the whole 34-year period from observations and AnEn. The yellow colour band shows the 10-90% range of the AnEn. The linear interpolation approach is shown in blue, part of which has been cut from the plot due to large differences in scale.

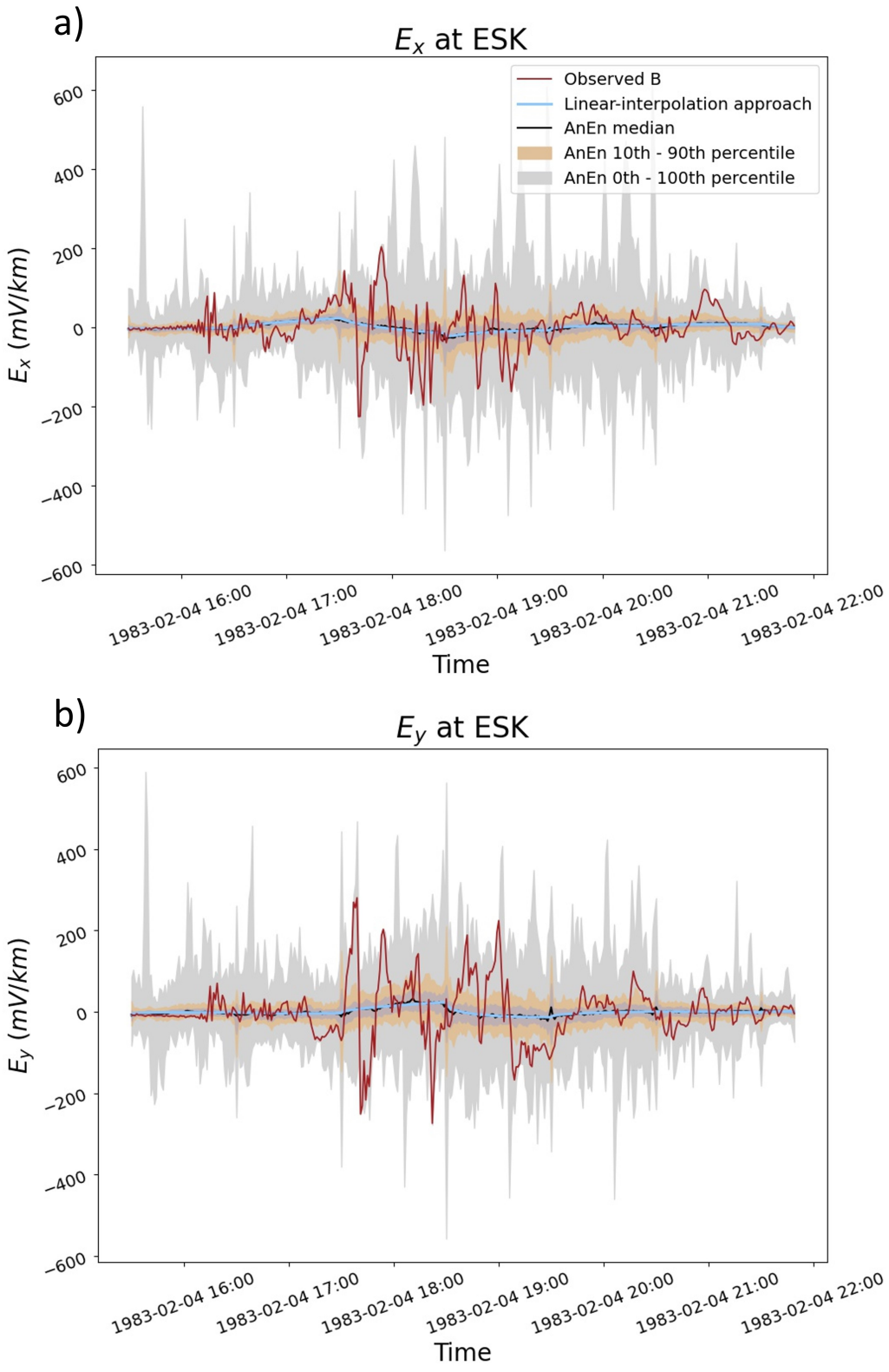


Figure 5: A six-hour time series from 1983-02-04 at ESK of the geoelectric field computed from the magnetic field using the MT-transfer function. The data is in the x (east-west) and y (north-south) directions in the geographic coordinate system. The red line shows the time series computed from the 1-minute observed time series, the colour bands show the spread of the geoelectric field computed from the analogue ensemble with the median in black, and the blue line shows geoelectric field computed from the linear-interpolation approach magnetic field.

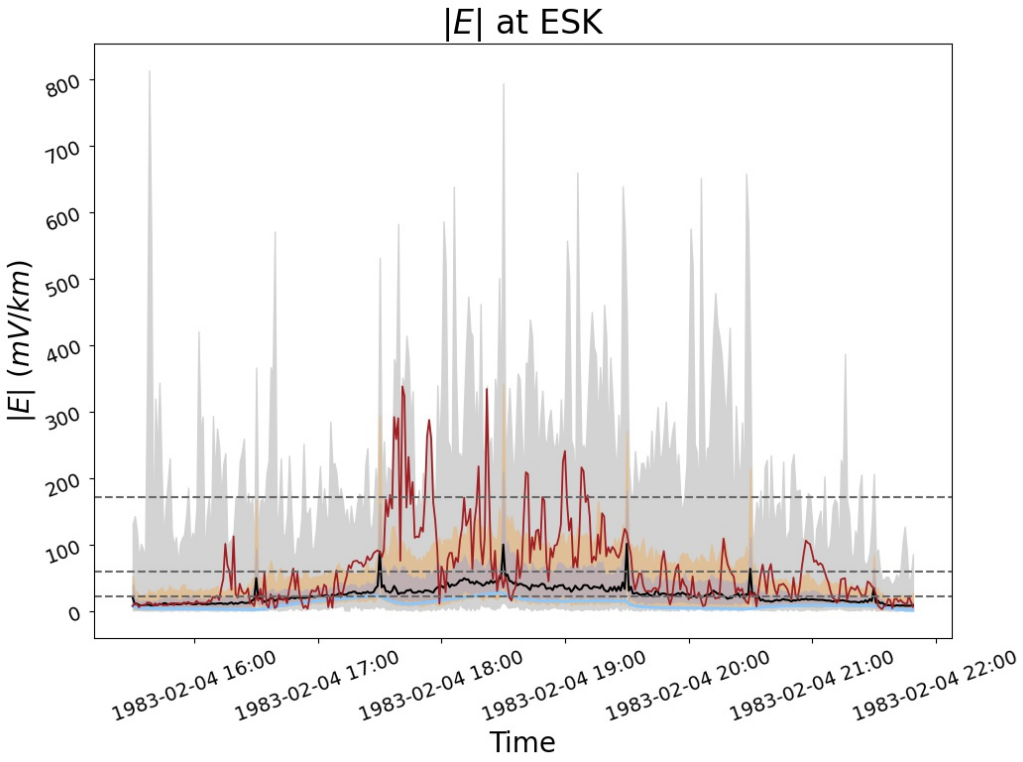


Figure 6: A six-hour time series from 1983-02-04 at ESK of the total magnitude of the geoelectric field computed from the magnetic field using the MT-transfer function. The red line shows the magnitude of time series computed from the 1-minute observed time series, the colour bands show the spread of the magnitude of geoelectric field computed from the analogue ensemble with the median in black, and the blue line shows the magnitude of geoelectric field computed from the linear-interpolation approach magnetic field.

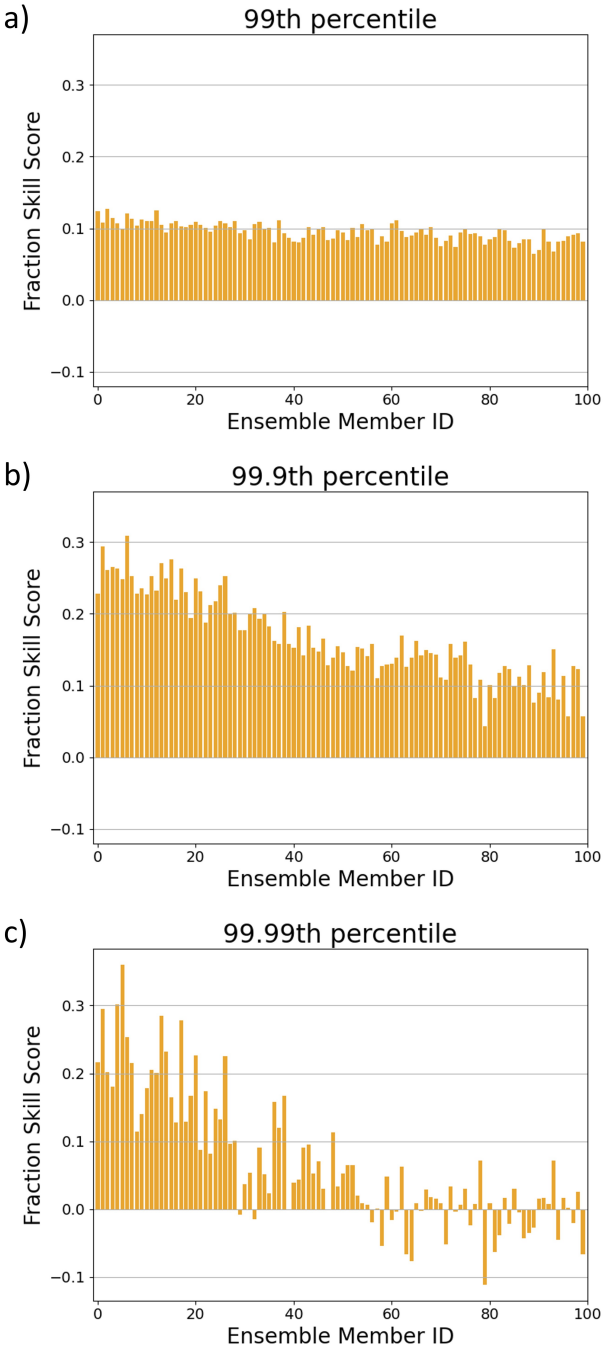


Figure 7: The fraction skill score (FSS) for each ensemble member. Ensemble members are ordered from best to worst analogues considered. A FSS of 1 represents a perfect model FSS of 0 represents a model with no skill over the reference. The time window for computing FSS is 60-minutes. a), b) and c) show FSS for events over the $99th$, $99.9th$ and $99.99th$ percentiles of the geoelectric field.

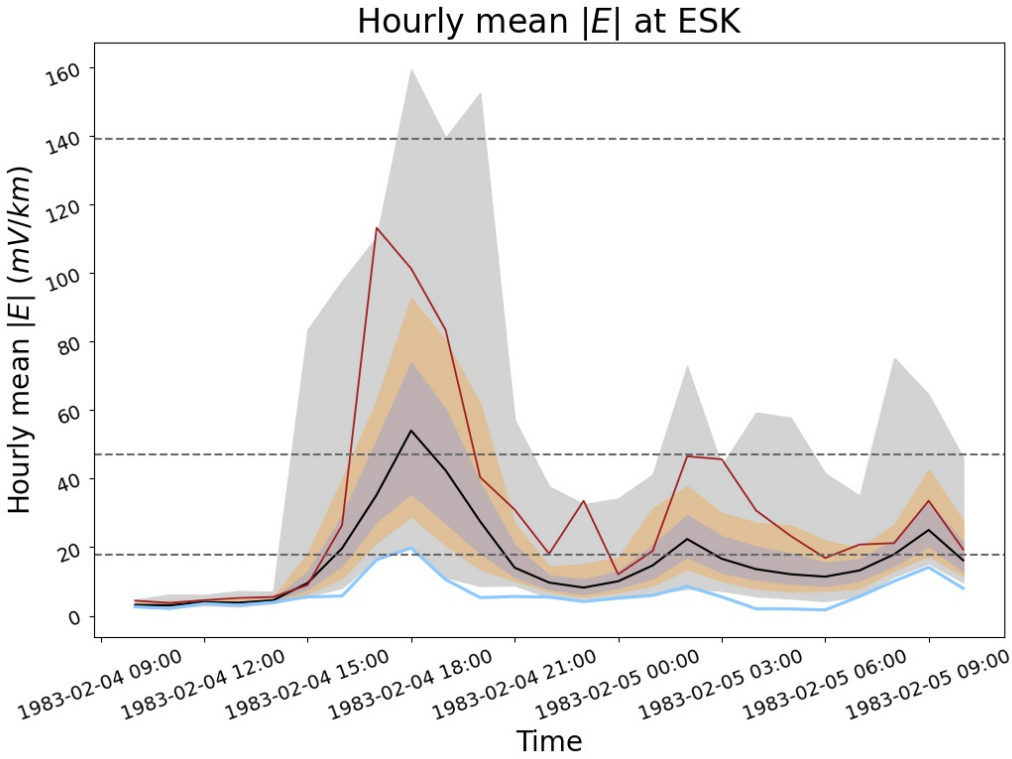


Figure 8: A time series from 1983-02-04 to 1982-02-05 at ESK of the 1-hour box-car mean of the magnitude of the geoelectric field computed from the magnetic field using the MT-transfer function. The red line shows the 1-hour mean of the magnitude of time series computed from the 1-minute observed time series, the colour bands show the spread of computed from the analogue ensemble with the median in black, and the blue line shows the 1-hour mean of the magnitude of geoelectric field computed from the linear-interpolation approach magnetic field.

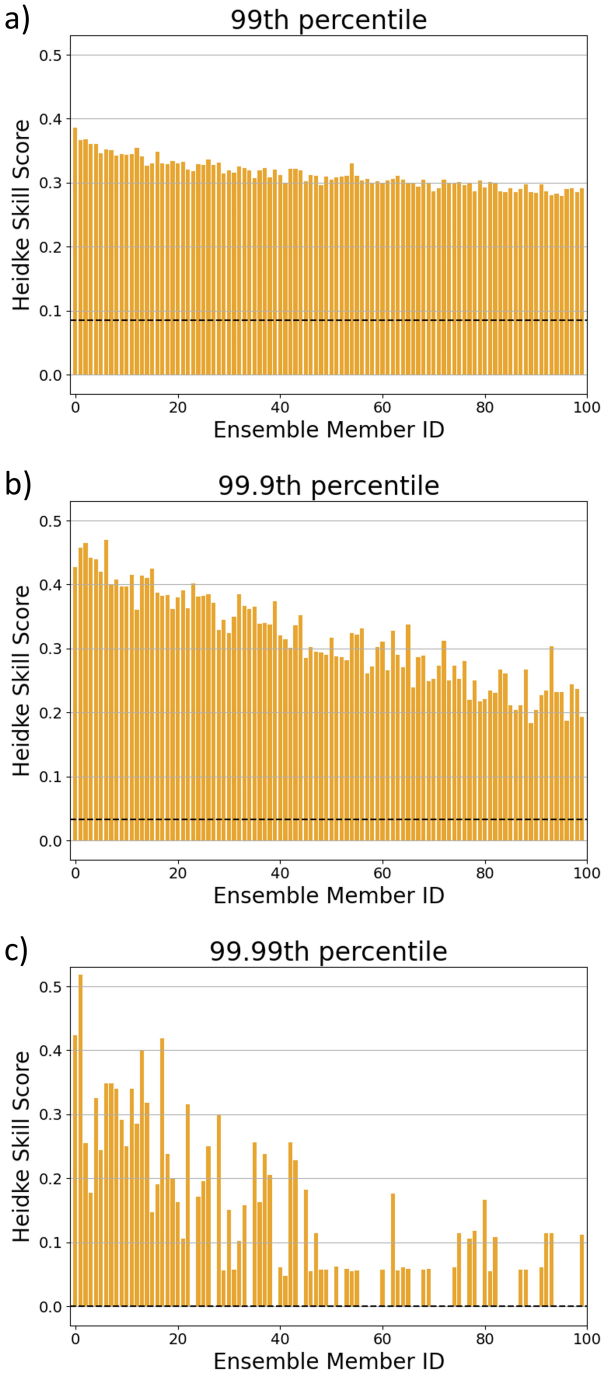


Figure 9: The Heidke skill score (HSS) for the three event thresholds on applied to 1-hourly $|\mathbf{E}|$ data. Ensemble members are ordered from best to worst analogues considered. A perfect forecast has a score of 1, a forecast with no skill over random prediction has a score of 0, and a forecast with every prediction incorrect has a score of -1. HSS is shown for each ensemble member. The black dashed horizontal line represents the HSS achieved by the linear-interpolation approach for each event threshold

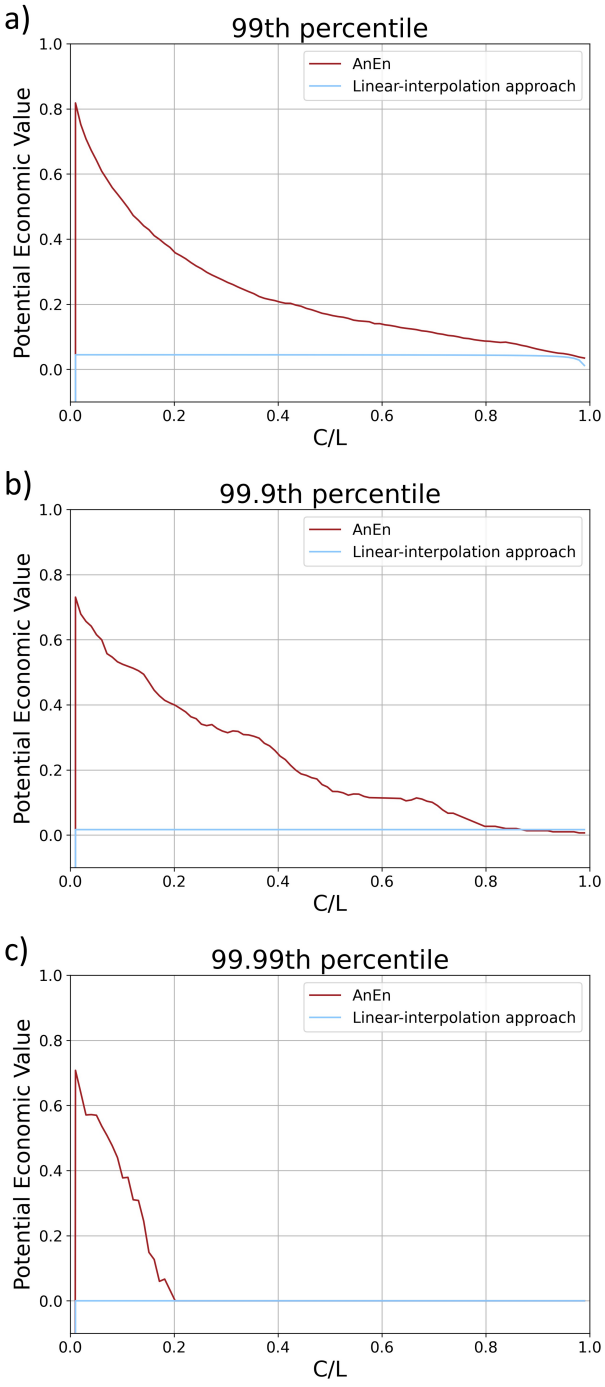


Figure 10: A Cost/Loss analysis showing the potential economic value (PEV) of the probabilistic AnEn downscaling method with respect to the undownscaled (linear-interpolation) reference method. A score of $PEV = 1$ represents a perfect forecast and $PEV = 0$ represents no value with respect to the reference method. a), b) and c) show PEV for events over the $99th$, $99.9th$ and $99.99th$ percentiles of the geoelectric field.