# Evaluation and Comparison of Geomagnetic Activity Forecasts

John Williamson (johwil@bgs.ac.uk), Ellen Clarke, Sarah Reay and Gemma Richardson
British Geological Survey, Edinburgh, UK

*European Space Weather Week 15*
*Session 10 - Space Weather Operations & Services*

## Introduction

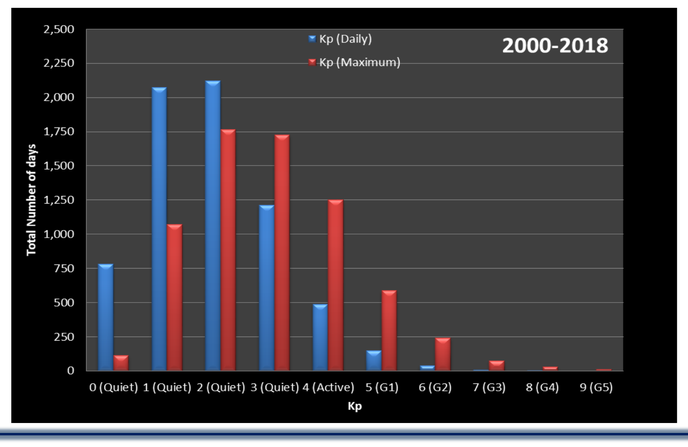
Geomagnetic activity forecasts of various types are provided by the British Geological Survey. These include categorical, human-derived forecasts of up to three days ahead as well as computer-derived time-series predictions of global and local daily (Ap and DRX). Users of these include institutes acting on behalf of government, space agencies concerned with thermospheric models of satellite drag, power companies interested in warning of possible geomagnetically induced currents, oil and gas companies involved in directional drilling and aurora borealis enthusiasts.

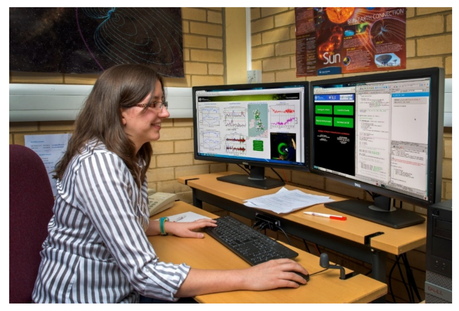
| Kp | BGS categories pre 2014 | | BGS categories since 2014 | | NOAA G-scales | |
|---|---|---|---|---|---|---|
| | Category | Description | Category | Description | Category | Description |
| <3+ | QUIET-UNSETTLED | Kp < 3+ | QUIET | Kp < 3+ | | |
| 3+ | ACTIVE | 3+ < Kp < 5- | ACTIVE | 3+ < Kp < 5- | | |
| 4- | | | | | | |
| 4o | | | | | | |
| 4+ | | | | | | |
| 5- | MINOR STORM | 5- < Kp < 6- | STORM G1 | 5- < Kp < 5+ | G1 | Kp = 5 |
| 5o | | | | | | |
| 5+ | | | | | | |
| 6- | MAJOR STORM | 6- < Kp < 8- | STORM G2 | 6- < Kp < 6+ | G2 | Kp = 6 |
| 6o | | | | | | |
| 6+ | | | | | | |
| 7- | | | STORM G3 | 7- < Kp < 7+ | G3 | Kp = 7 |
| 7o | | | | | | |
| 7+ | | | | | | |
| 8- | SEVERE STORM | Kp > 8- | STORM G4 | 8- < Kp < 9- | G4 | Kp = 8 |
| 8o | | | | | | |
| 8+ | | | | | | |
| 9- | | | | | | |
| 9o | | | STORM G5 | Kp = 9o | G5 | Kp = 9 |

Previously the forecasts derived by the team were based on a 4-category classification. From 20th May 2014 onwards a 7-category classification was adopted, based on the NOAA G-Scale, which in turn is based on Kp levels. The diagram (left) shows these levels and how they relate before and after the change made. In a previous analysis [**1**] forecasting performance was investigated from 2000 to 2013. This work is extended to 2018 and new comparisons with ARIMA based [**2**] predictions made.

The observed distribution of geomagnetic activity as determined by the Kp index (daily and maximum) over the period of forecast evaluation (2000 to 2018) is shown (right).

Quiet periods greatly outnumber Storms, making forecasting these events notoriously difficult. This skewed distribution can be accounted for in forecast evaluation by using equitable skill scores.



A forecaster (left) making 1, 2 and 3-day ahead forecasts. Various solar and solar wind observations, data and models available in the public domain, as well as in-house products are analysed and interpreted.

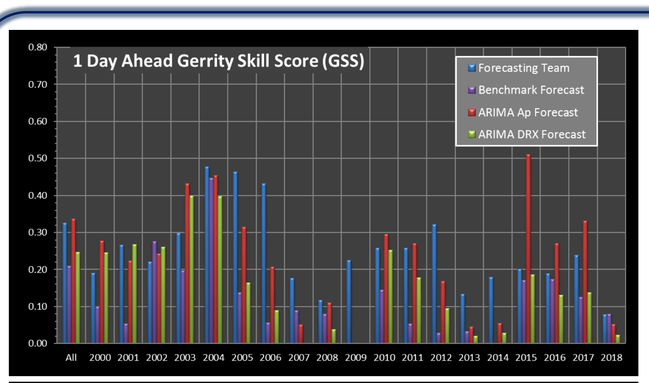
The forecast made by the duty forecaster 7th Sep 2017 is shown (right). The storm, which was predicted due to a CME associated with the ~X10 solar flare of the 6th Sep, peaked at G4 and averaged G2.

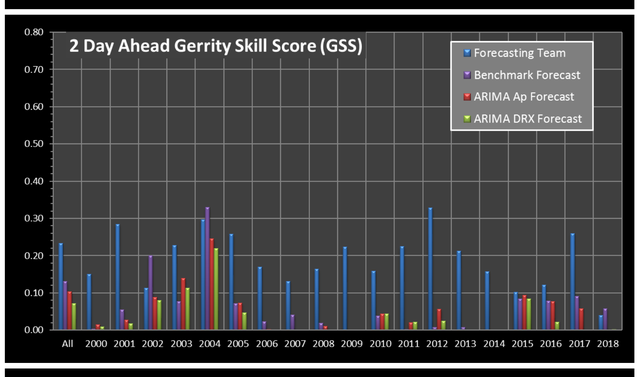
| Forecast period (noon-to-noon GMT) | Forecast Global Activity level | |
|---|---|---|
| | Average | Max |
| 7 SEP-8 SEP | ACTIVE | STORM G3 |
| 8 SEP-9 SEP | STORM G2 | STORM G3 |
| 9 SEP-10 SEP | ACTIVE | STORM G2 |

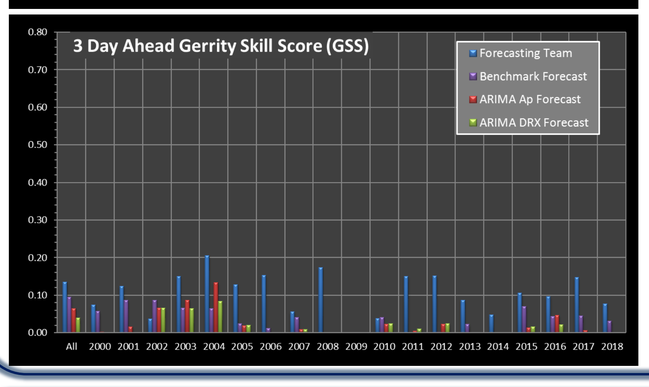
The forecast as it would have looked using the `old' 4-category system is shown for comparison. The introduction of likely maximum levels was intended to provide additional information to the users and to help the forecaster with their decisions.

| Forecast period (noon-to-noon GMT) | Forecast Global Activity level |
|---|---|
| 7 SEP-8 SEP | ACTIVE |
| 8 SEP-9 SEP | MAJOR STORM |
| 9 SEP-10 SEP | ACTIVE |

## Results 1




The skill scores of the forecast team are compared with those derived from three other methods over the same time periods.

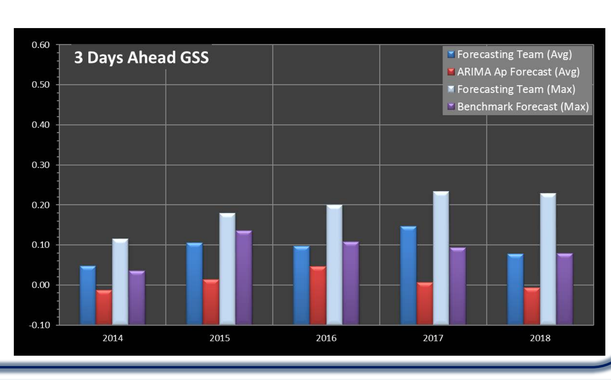
1) A benchmark of persistence and recurrence (average of 1 and 27 days ago)
2) Planetary activity as predicted by the ARIMA method for Ap **[2]**
3) Local activity (Northern UK) as predicted by the ARIMA method for DRX (Daily average of the 24 hourly ranges in the *X* component) at Lerwick.

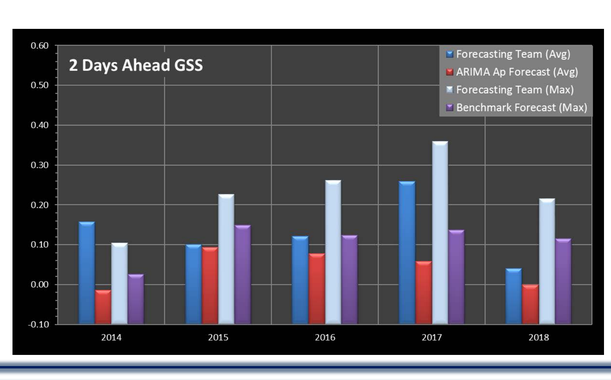
The Gerrity Skill Score (Equitable Threat Score) **[3]** is used to compare forecasts. A higher score indicates more skill. Contingency matrices for each year and all years (example on right) for each forecast method are derived. 4x4 scoring (S) matrices (an example of the 7x7 version for forecasts from 2014 onwards is shown in Results 2) were derived and applied to ensure the scores are equitable.

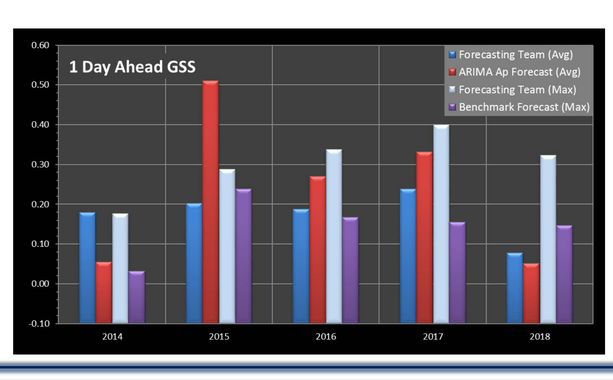
Average noon-noon forecasts for all methods were compared over 18 years (left) incorporating both old and new version of the forecast. G2-G5 storms in the new category (since 2014) were all classed as Major-Storm for backward compatibility with the historical forecasts.

Here the S matrices are derived from the year being analysed as opposed to the full 18 years of data. This reduces further any solar cycle dependence as previously reported **[1]**.

| Geomag Forecasters Forecast: 4 x 4 contingency matrix (E) | | Forecast Category | | | | Marginal Total |
|---|---|---|---|---|---|---|
| | | Q-U | ACTIVE | MINOR | MAJOR | |
| Observed Category | Q-U | 3498 | 511 | 37 | 2 | 4048 |
| | ACTIVE | 283 | 291 | 43 | 6 | 623 |
| | MINOR | 42 | 89 | 31 | 4 | 166 |
| | MAJOR | 9 | 37 | 24 | 16 | 86 |
| Marginal Total | | 3832 | 928 | 135 | 28 | 4923 |



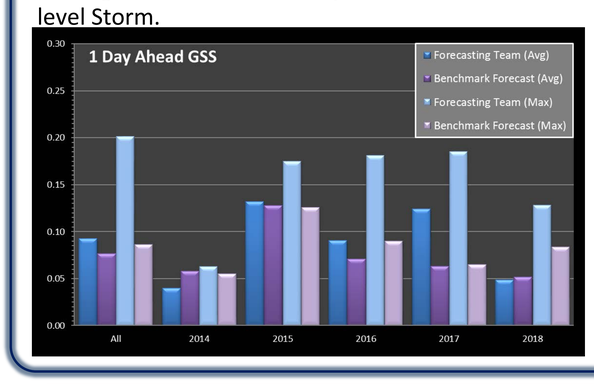

Since 2014 the maximum 3-hour activity in a 24-hour noon-noon period is also forecast. The skill scores for these are compared to those for the maximum forecast by the benchmark method and plotted (below) alongside the results for the daily average and ARIMA method for 2014-2018.

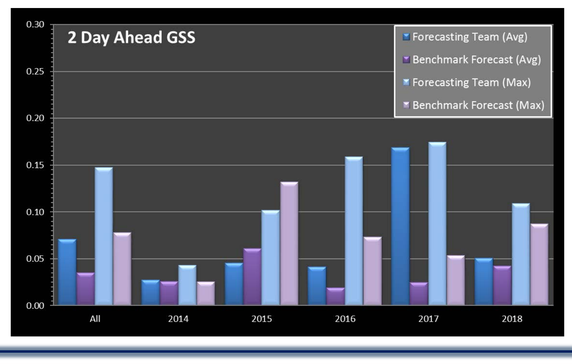
## Results 2

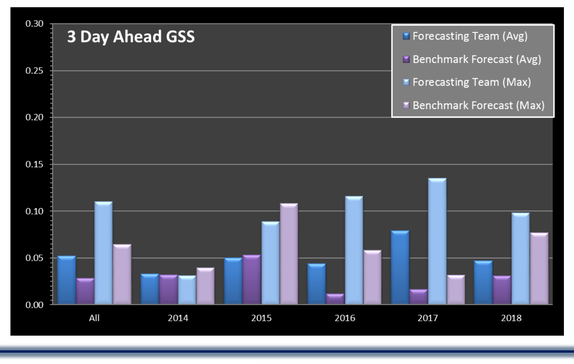
Evaluation of the new forecasting classification scheme is possible for years 2014-2018. 7x7 contingency tables and Gerrity reward-penalty scoring matrix (right) were derived and used to compute GSS (below).

The high score for 2017 2-day ahead forecast was due to the correct forecast of a G2 level Storm.

| 7 x 7 reward-penalty or scoring matrix (S) as per Gerrity (1992) | | Forecast Category | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | QUIET | ACTIVE | STORM G1 | STORM G2 | STORM G3 | STORM G4 | STORM G5 |
| Observed Category | QUIET | 0.07 | -0.16 | -0.33 | -0.50 | -0.67 | -0.83 | -1.00 |
| | ACTIVE | -0.16 | 0.44 | 0.27 | 0.10 | -0.06 | -0.23 | -0.40 |
| | STORM G1 | -0.33 | 0.27 | 5.86 | 5.69 | 5.52 | 5.36 | 5.19 |
| | STORM G2 | -0.50 | 0.10 | 5.69 | 23.87 | 23.70 | 23.53 | 23.37 |
| | STORM G3 | -0.67 | -0.06 | 5.52 | 23.70 | 69.51 | 69.34 | 69.17 |
| | STORM G4 | -0.83 | -0.23 | 5.36 | 23.53 | 69.34 | 140.91 | 140.75 |
| | STORM G5 | -1.00 | -0.40 | 5.19 | 23.37 | 69.17 | 140.75 | 228.84 |

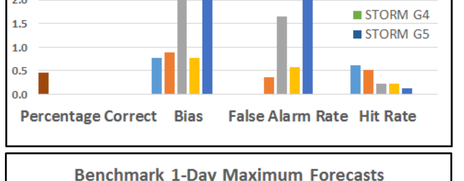


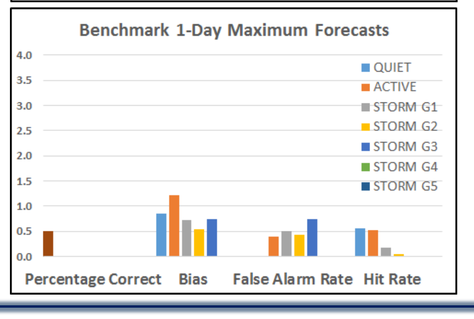




Other forecast verification statistics - Bias, Percentage Correct, False Alarm Rate and Probability of Detection or Hit Rate – have been derived for the maximum forecasts and compared against the benchmark (right).

Whilst the team out-perform the benchmark on hit rate, it comes at the expense of higher false alarm rates and high bias scores.




## Summary and Further Work

Results show that forecasts made by the forecasting team compare well with other methods, however there are times when they are out-performed. Most notably 2015-2017 when the ARIMA method attains higher skill scores in the 1-day ahead daily average forecasts. This is likely due to the high level of persistent activity during the declining phase of the solar cycle.

For 2- and 3-day ahead the team clearly demonstrate more skill than the benchmark or ARIMA methods.

Higher skill scores are achieved for forecasts of a maximum than a daily average. This is reasonable as a forecaster has eight 3-hour chances to get a correct activity level, compared to one opportunity for the average to be correct.

Further comparisons of the maximum forecasts with the operational 3-hour ARIMA forecasts of ap is now possible. This will be included in future work.

## Acknowledgments

The authors would like to thank all BGS staff past and present who have contributed to the Forecasting Team including Orsi Baillie, Thomas Humphries, Brian Hamilton, Alan Thomson and Guanren Wang. We also thank Peter Stevenson for his help with debugging software.

The definitive Kp and Ap indices are derived and made available at GFZ, Potsdam, Germany on behalf of the ISGI.

## References

[1] Clarke, E. and Thomson, A.W.P. Forecast Evaluation as Applied to Geomagnetic Activity Categories (2013), Conference Presentation, European Space Weather Week 10. www.stce.be/esww10/contributions/public/talks/Session12/05-ClarkeEllen/Clarke_ForecastVerification.pdf )
[2] Thomson, A.W.P., Clark, T.D.G. and Kerridge, D.J. (1992) Computer Algorithms and FORTRAN Programs for Forecasting Solar and Geomagnetic Activity in the Short-term, British Geological Survey Technical Report WM/92/19C
[3] Gerrity, J.P.Jnr. (1992) A Note on Gandin and Murphy's Equitable Skill Score. Mon. Wea. Rev., 120, 2709-2712.