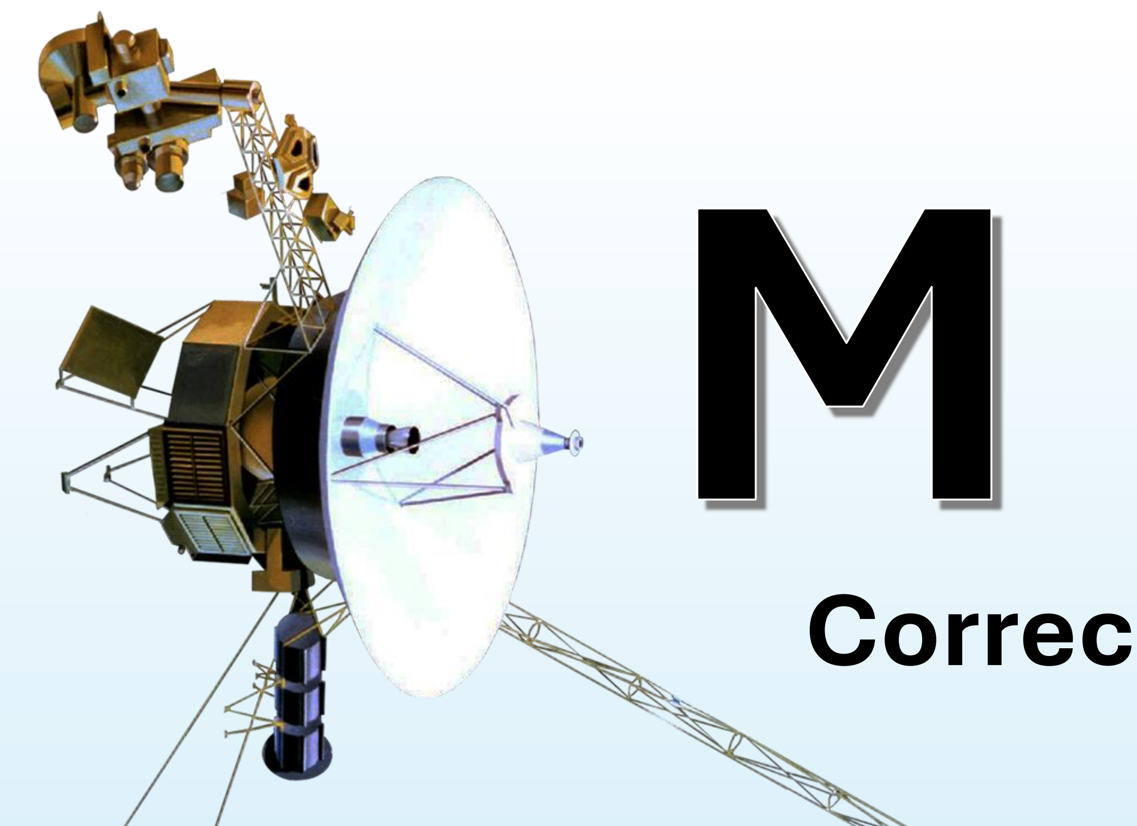
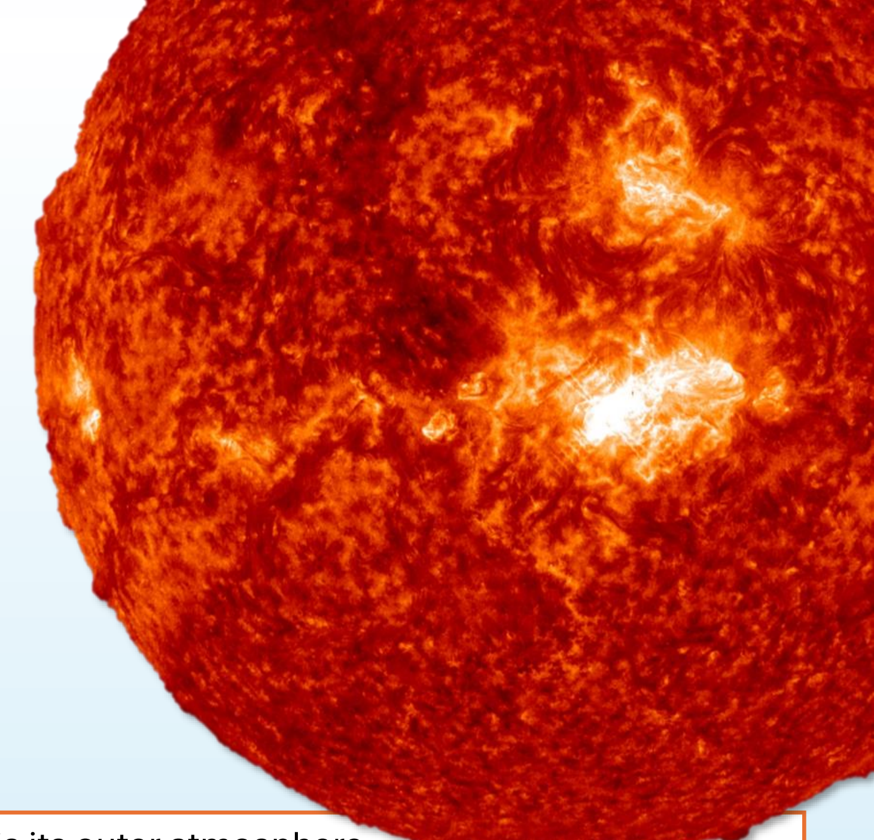


MIND THE GAP



Correcting gap-induced bias from sparsely-sampled time series of solar wind turbulence

VICTORIA UNIVERSITY OF WELLINGTON TE HERENGA WAKA

Daniel T. S. Wrench & Tulasi N. Parashar

daniel.wrench@vuw.ac.nz School of Chemical & Physical Sciences

Voyager. Twin spacecraft launched in 1977 that continue to send back data, albeit increasingly incomplete. Being able to make the most of these measurements is crucial to understanding turbulent dynamics in the outer heliosphere and interstellar medium.

The Sun. Why is its outer atmosphere so much hotter than its surface? How is energy dissipated in a fluid with practically zero collisions such as the solar wind? Can we better predict space weather? Improving our understanding of turbulence could help answer these questions.

THE PHYSICS

The solar wind is a plasma that continuously flows out of the Sun (Fig. 1). During its supersonic propagation through the solar system, it becomes **turbulent**. Understanding this turbulence is key to improving models of energy dissipation in collisionless plasmas, particle transport in the heliosphere, and, in turn, our forecasts of space weather.

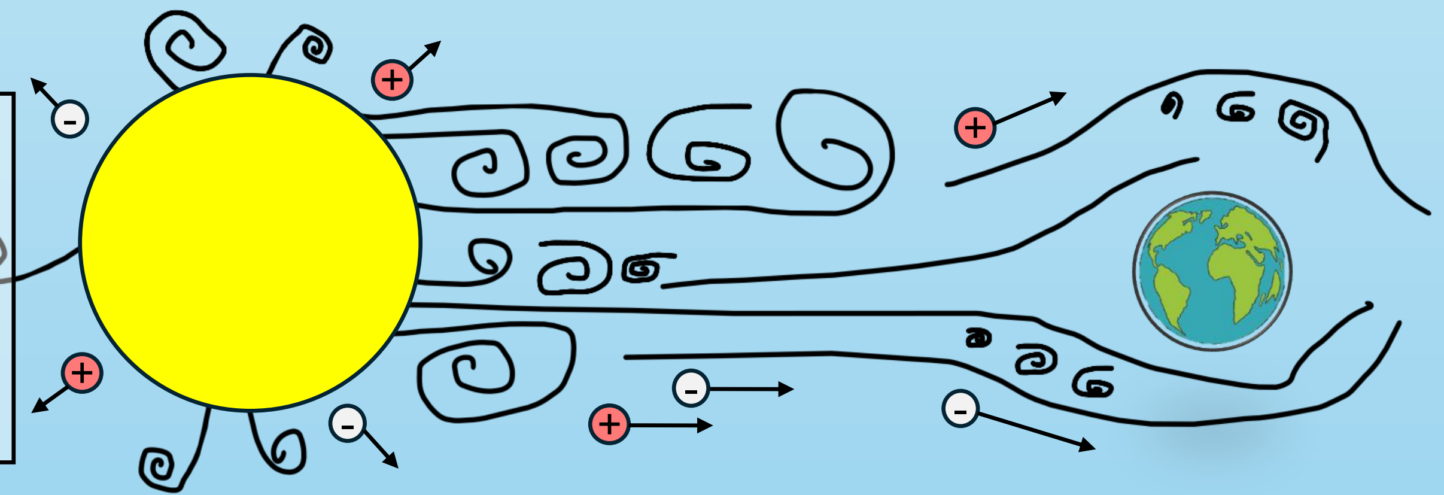


Figure 1: An artistic representation of the solar wind, consisting of protons and electrons, as it forms turbulent eddies and is deflected around the Earth's magnetosphere. *Not to scale.*

THE STATISTICS

Due to their highly unpredictable and multi-scale nature, turbulent flows are typically studied using statistics that quantify the distribution of energy across scales. One such statistic is the **structure function (SF)**. This gives the moments of the distributions of increments, $\mathcal{P}[x(t+\tau)-x(t)]$, as a function of lag τ. The slope of the SF in log-log space is computed and compared with theoretical predictions. The 2nd-order ($p=2$) SF is called the **variogram** in geostatistics (Fig. 2). The 4th-order SF (scale-dependent kurtosis) is used to study **intermittency**.

$$S_p(\tau)=\langle|x(t+\tau)-x(t)|^p\rangle$$

Equation 1: Formula for calculating the p^{th}-order structure function (SF). Angle brackets denote an ensemble average. In geostatistics, this is referred to as Matheron's *method-of-moments* estimator, and $S_2(\tau)$ is known as the variogram.

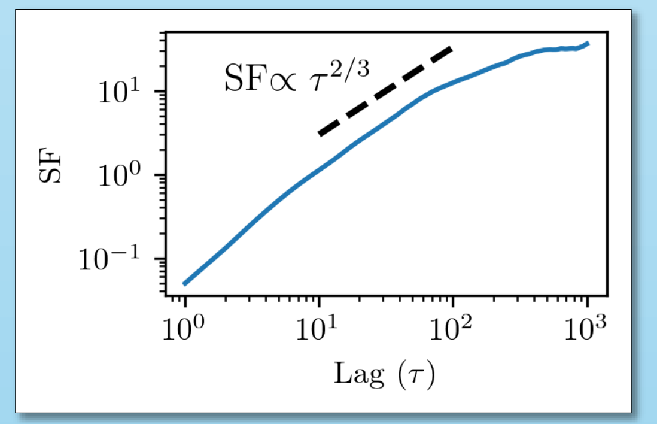


Figure 2: Example second-order SF with expected K41 power-law.

THE PROBLEM

Unfortunately, while often comprising many thousands of observations, intervals of the solar wind are commonly plagued by gaps. This is particularly a problem for data from the distant *Voyager* spacecraft. The SF can be computed from un-evenly sampled series, but it becomes increasingly distorted as gaps affect different lags to different extents. Being able to account for this distortion is key to maximising the amount of robust scientific output from such sparse time series.

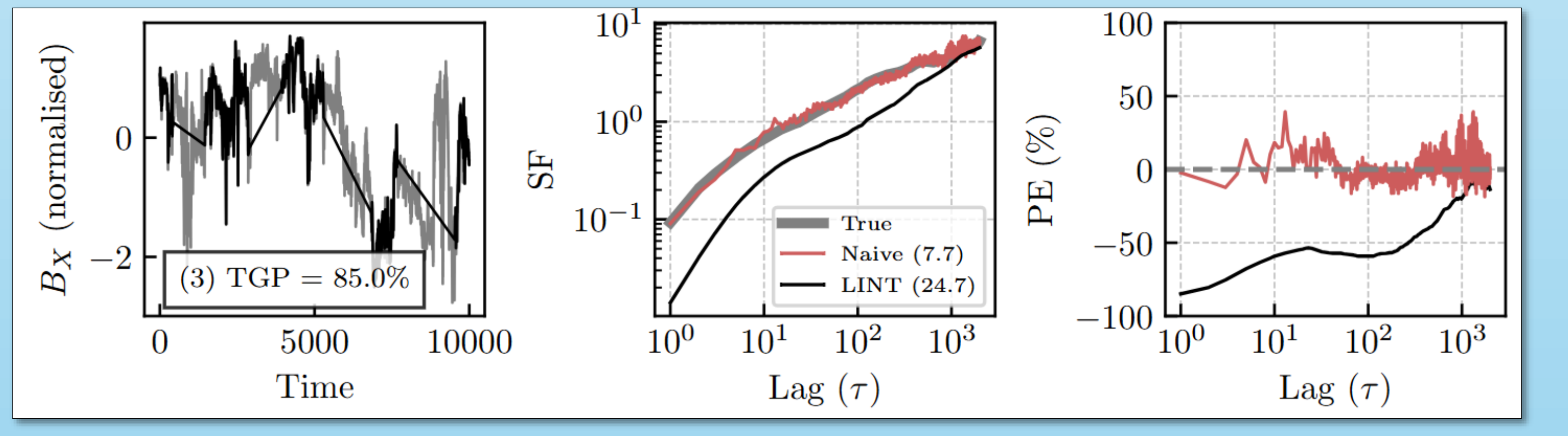


Figure 3: Case study on the effect of simulated gaps in magnetic field data from the Wind spacecraft (left panel) on estimates of the second-order structure function (middle panel). This interval has had 85% of data removed, mostly in contiguous chunks. Numbers in brackets give the *mean absolute percentage error* of naïve (no handling) and LINT (linear interpolation) methods, with percentage error as a function of lag given in the right panel.

METHOD

We simulate a variety of gap types for a large set of magnetic field intervals from the Parker Solar Probe spacecraft (see below). Using Eq. 1, we estimate the 2nd order SF from these intervals **(naïve estimator)**, as well as linearly-interpolated versions **(LINT estimator).** We calculate the percentage error (PE) at each lag, relative to the SF from the complete dataset.

We then average the LINT PE for each (binned) combination of missing fraction and lag (Fig. 4, bottom right). We then convert this average error into a multiplicative correction factor, which we apply to LINT estimates of the SF from the Wind spacecraft, as a test of our correction procedure.

Parker Solar Probe
0.05 AU from the Sun

TRAINING SET
10,731 *PSP* intervals, standardised and randomly gapped in 25 different ways with up to 95% missing data, in both contiguous chunks and individually missing points

Wind
1 AU from the Sun

TEST SET
660 *Wind* intervals, processed as above

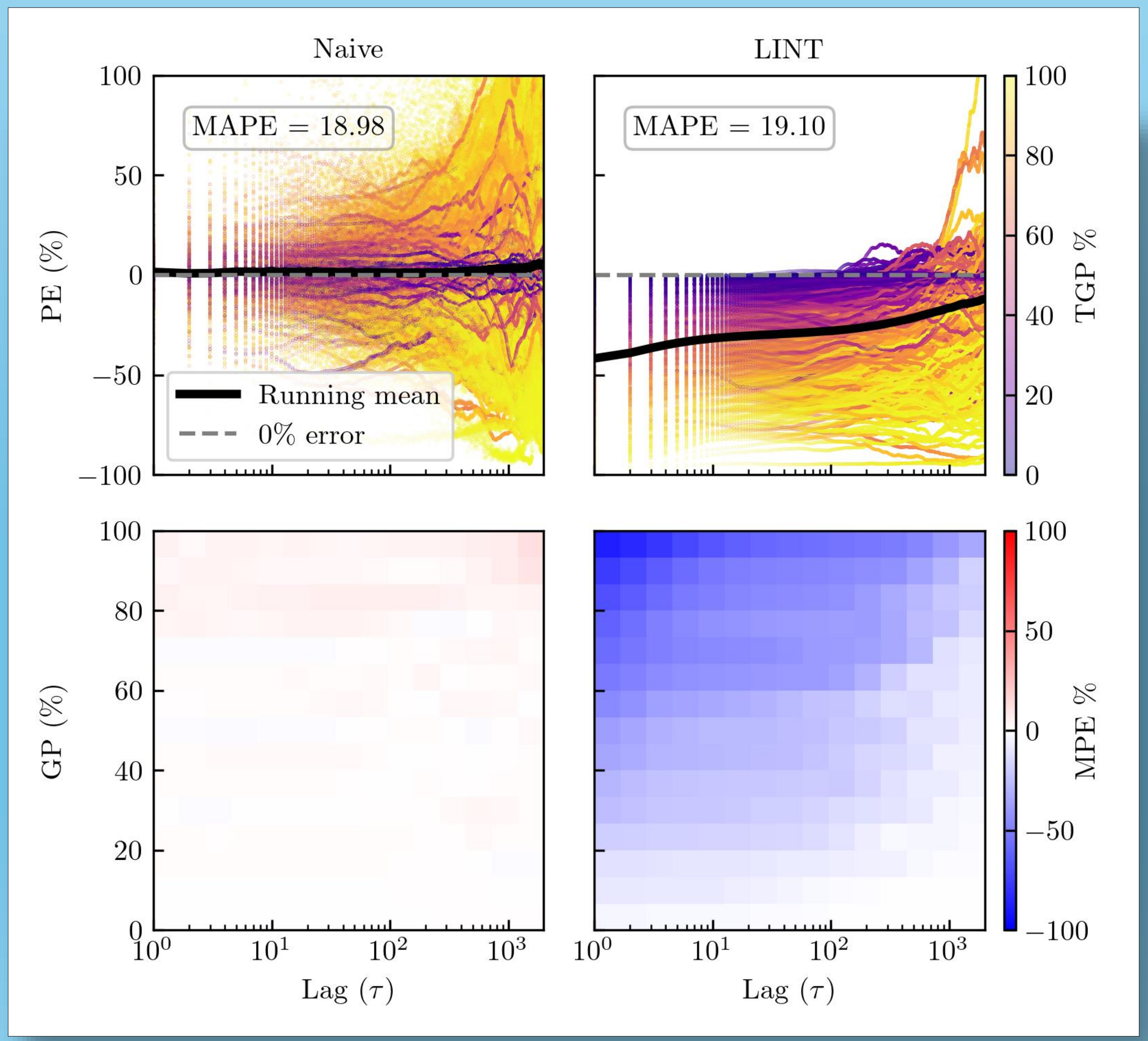


Figure 4: Simulation results of SF error caused by gaps. *Top row:* percentage error trendlines for each artificially-gapped PSP interval, as in the right-hand panel of Fig. 1, coloured according to its total gap percentage. *Bottom row:* mean percentage error for each bin of lag-specific gap percentage (GP) and lag. The consistent underestimation of the LINT estimator is clear, so this is used as the basis for our correction factor.

RESULTS

As shown in the case study in Fig. 3, no interpolation of the gaps (naïve method) leads to a very noisy SF, with both negative and positive errors. However, Fig. 4 (top left) shows that the average error across lags is approximately zero. Meanwhile, the smoothing effect of linear interpolation (LINT) causes systematic underestimation, with a clear dependence on both lag and percentage missing.

We find that this bias in the LINT estimate is consistent enough that we can use it to partially "de-bias" SF estimates and ultimately reduce the overall error of the curve, relative to the other two methods, for sparsity > 25%. We demonstrate this using a correction factor learnt from PSP data but applied to Wind intervals, as shown in Fig. 5 and statistically evaluated in Fig. 6.

This then gives confidence for us to apply this correction factor to structure function estimates for which we have no "ground truth", such as Voyager intervals from the local interstellar medium, which can have up to 80% of data missing.

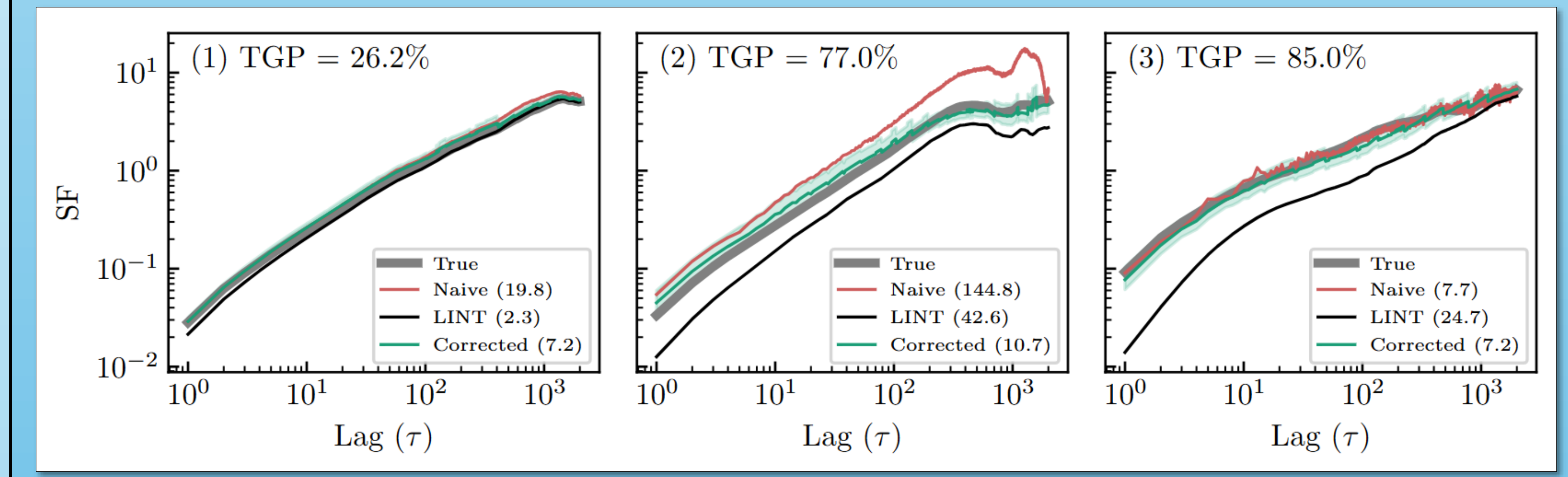


Figure 5: Case studies of applying the correction factor to three Wind intervals, including the interval in Fig. 1. MAPEs are given in brackets. An estimate of uncertainty in the corrected interval is based on the spread of errors in each bin and is shown by the shaded green region, indicating plus and minus two standard deviations.

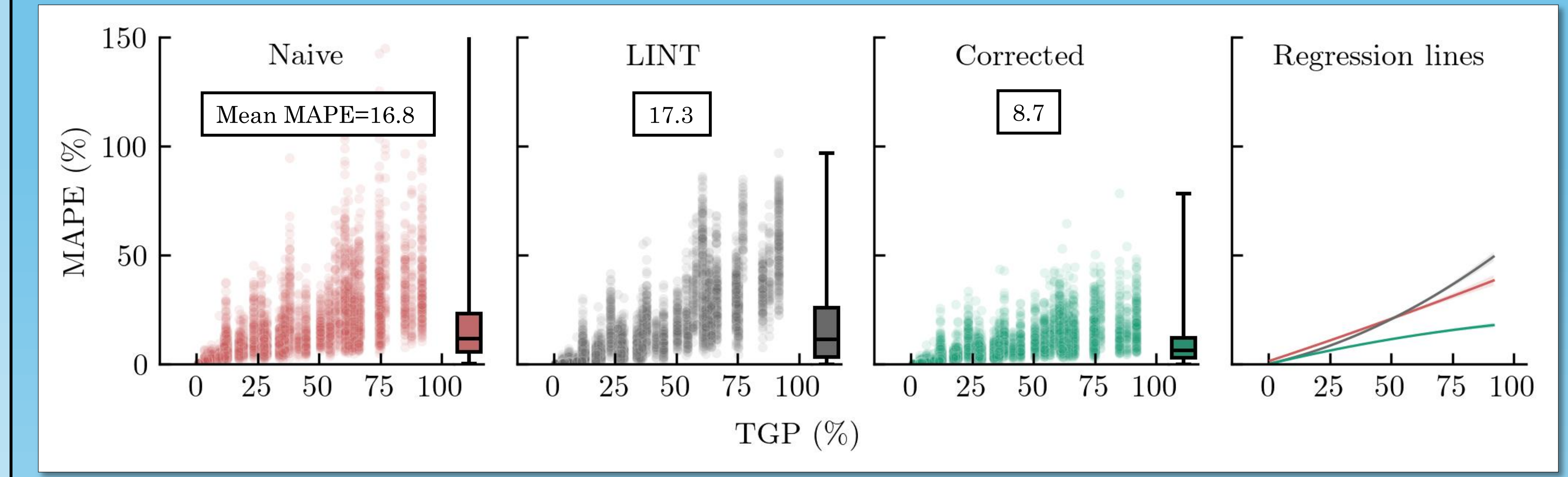


Figure 6: Error as a function of lag for each of the three structure function estimation methods for the Wind test set. On the right axis of each scatter plot is a box plot showing the marginal distribution of errors: The final panel shows order-2 polynomial regression lines fitted to each scatterplot.