3. Model Documentation and Write-Up

1. My name is Tom, and I'm currently a physics PhD student at the University of Oxford. My research is in physical oceanography, in particular Gulf Stream turbulence. My day-to-day work involves statistical analysis of observational datasets and model output, as well as some theoretical approaches. In my spare time I'm a keen cyclist.

2. My method involved a mix of persistence, auto-regressive models, linear regressions and exponential models. As the challenge involved multiple time series, the particular method applied depended on my judgement, after manual inspection of the data. Tailoring the approach for each time-series individually proved successful as opposed for a single technique across all of them. The behaviour of the time-series varied massively between sites and species. One time-series might show a linear increase due to increased food availability, where a neighbouring site may exhibit exponential decrease due to humans entering the region; with so many potential factors affect the inter-annual variability of nest counts, it made the challenge particularly difficult.

3. For one particular attempt, I used the ERA-interim reanalysis product: a global reconstruction of the state of the climate using a combination of observations and models. I extracted quantities such as air temperature, winds, sea surface temperature at all sites across Antarctica. The aim was to incorporate these variables into a model, in the hope that changes in the environmental correlated with changes in the nest counts. It was a physically motivated model, which was appealing, but it unfortunately did not improve the prediction accuracy for this competition.

4. A significant amount of time was spent plotting each of the 600+ time series in the training data, and manually inspecting each one. This was an important part of the challenge as you gain an intuition of how each time series behaves, and what models might be appropriate for that particular site.

5. I conducted some cross-validation by using the most recent value of each time-series in the training data as the test data, and computing the AMAPE score for predictions of that value. This was how I discovered that applying one method across all time series always performed worse that simple persistence. This guided me towards more “personalized” methods for each times series.

6. One feature that could have been potentially useful is whether humans live at or near the site in question. Some sites are undisturbed while others have human settlements, which can affect the local penguin populations. Another feature would be estimates of birth/death rates, as you were unable to differentiate whether a change in nest counts were due to deaths/births, or due to migration in or out of that site.

7. No.

8. I produced one graph showing how linear trends in population counts, related to linear trends in the air temperature. This can be found in my submitted report.

9. The day before the submission deadline I went to a conference where one of the speakers used Network Auto-Regressive (NAR) models to models how disease spread across different counties in the UK. This type of model would have be perfect for modelling interactions between different sites or regions in Antarctica. If I were to continue working on this challenge, I would definitely explore some simple NAR models.

10. Yes.