TAF performance

30 hour TAFs (South) - November 2022

TAF Verification using the Gerrity Skill Score

Since 2005 the TAF verification scheme used by the Met Office has been based on a Service Quality Index (SQI). The SQI has served the Met Office well, however from April 2020 a new verification scheme has been endorsed by the CAA and adopted by the Met Office. This scheme is based on a statistical measure called the Gerrity Skill Score (GSS).

What is the Gerrity Skill Score (GSS)?

The GSS is a measure of forecast discrimination. This means that it assesses whether a TAF category that was forecast actually occurred. For example, a good score for the highest visibility category is obtained if it regularly occurred whenever it was forecast, but rarely occurred when it was not forecast.

The GSS answers the question *"What was the accuracy of the forecast at predicting the correct category, relative to that of random chance?"* and ranges from -1 to 1, with 0 indicating no skill and 1 indicating that every occurring category has been correctly predicted.The GSS gives a greater reward to meteorologists who *"stick their neck out"* in an attempt to correctly predict rare events, rather than those who adopt a more conservative approach.

Unlike the SQI more than one category threshold is used by the GSS (*see table 1*). The Met Office assess all the TAF cloud base and visibility categories defined by ICAO. At the time of writing, the following categories are used for TAF assessment:

Surface visibility	Cloud base
10km or more	1500ft or more
5000m to 9999m	1000ft to 1499ft
1500m to 4999m	500ft to 999ft
800m to 1499m	201ft to 499ft
350m to 799m	200ft or less
349m or less	

Table 1 - Categories used in TAF verification

An assessment of forecast accuracy

The GSS does not only consider whether the observed category has been forecast correctly, it also considers the magnitude of any forecast error. Categories that often occur, and should therefore be easier to forecast, are harshly penalised by the GSS if they are incorrectly predicted. On the other hand, categories that rarely occur, and are therefore difficult to forecast correctly, are only penalised less if they are incorrectly predicted (if the event is particularly rare, a small forecast error can lead to a small reward).

A GSS of 0.5 indicates that, over the 3-year period of analysis, the issued TAFs were twice as accurate as random forecasts that have been based on how frequently each category was observed. In a Gerrity Score analysis of accuracy, greater credit is given to correct (and almost correct) forecasts of rare events and less credit is awarded to correct forecasts of common events.

Effects of climatology

The GSS can be sensitive to changes in the climatology. If a very rare event is correctly predicted it will provide a large boost to the score. However, when that event "drops out" of the rolling 3-year calculation the score may suddenly seem to drop if no similar extreme has been correctly predicted in the meantime.

Improved interpretations of TAF change groups

The new verification scheme has tried to improve on the TAF term interpretations adopted by the SQI approach. The new interpretations used to assess each TAF term are based on the definitions within ICAO documentation, namely:

- TEMPO: the alternative category should occur at least once and up to half the time
- BECMG: a transition period when the alternative category should occur at least once
- PROB30/40: the alternative category is forecast with a probability of 30% / 40% at each METAR time
- PROB30/40 TEMPO: the alternative category is forecast with a probability of 30% / 40% at least once and for up to half the time

What is the Peirce Skill Score (PSS)?

The PSS is an additional analysis of forecast accuracy that has been used to assess each cloud base and visibility category shown in *table 1*.

PSS graphs are calculated for each individual cloud base and visibility category to evaluate how accurately the forecasts are compared to the other forecast categories (benchmarked against random observed chance). For example, how accurately TAFs issued at Heathrow predict visibilities between 350m and 799m after accounting for being correct by random forecast chance and benchmarked against the expected accuracy achievable by random observed chance.

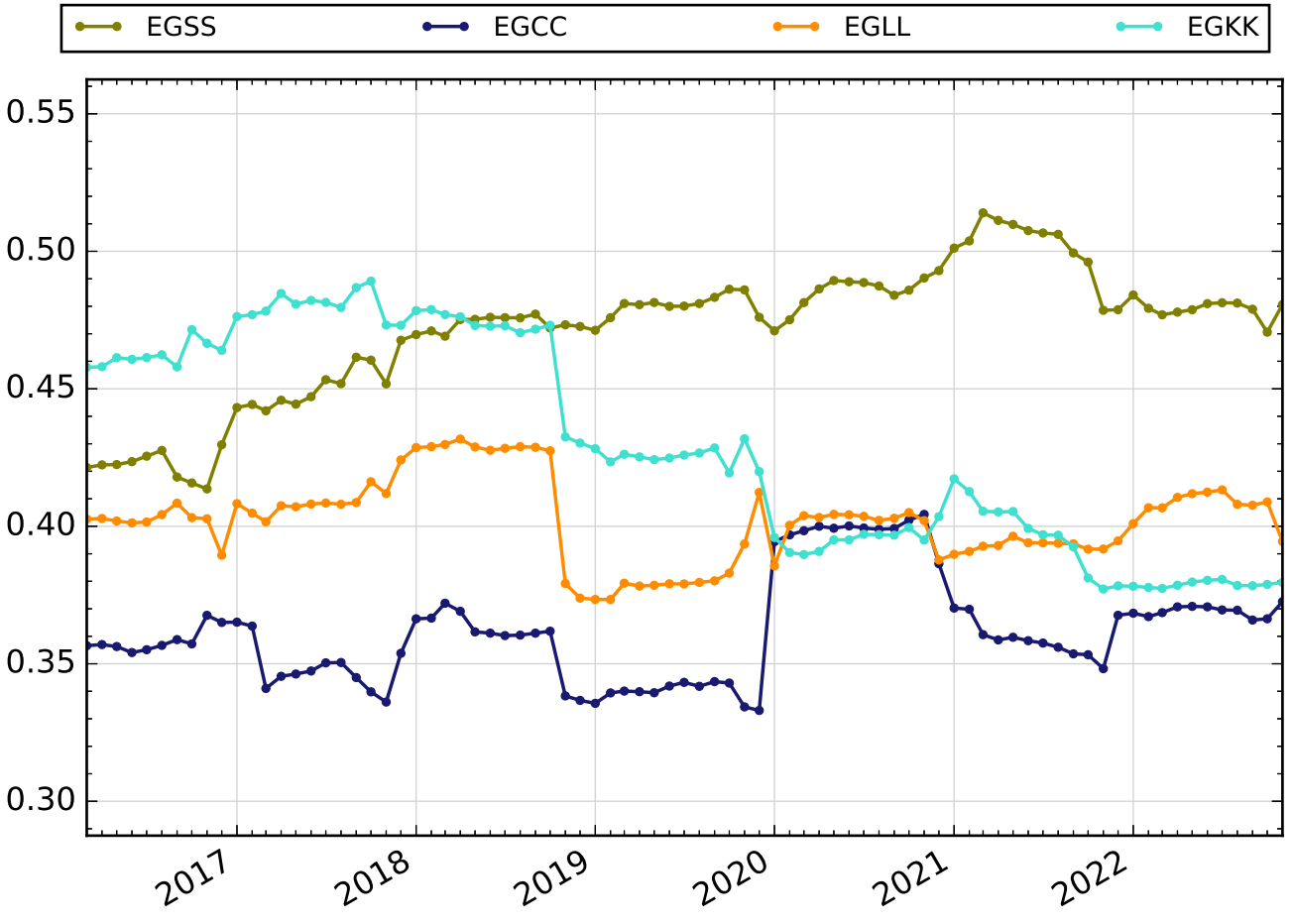
The obtained results are always between -1 and 1. A score of 0 means that, overall, the TAFs are no more skilled than randomly issuing forecast categories according to the climatology at the airport. A score of 1 means the forecast issued was always totally correct.

It is more difficult to forecast the weather at some airports than others and at the more difficult locations to forecast for the proportion correct will tend to be lower. The PSS graphs can be used to find out how much skill was added by the meteorologists for each forecast category and for each airport.

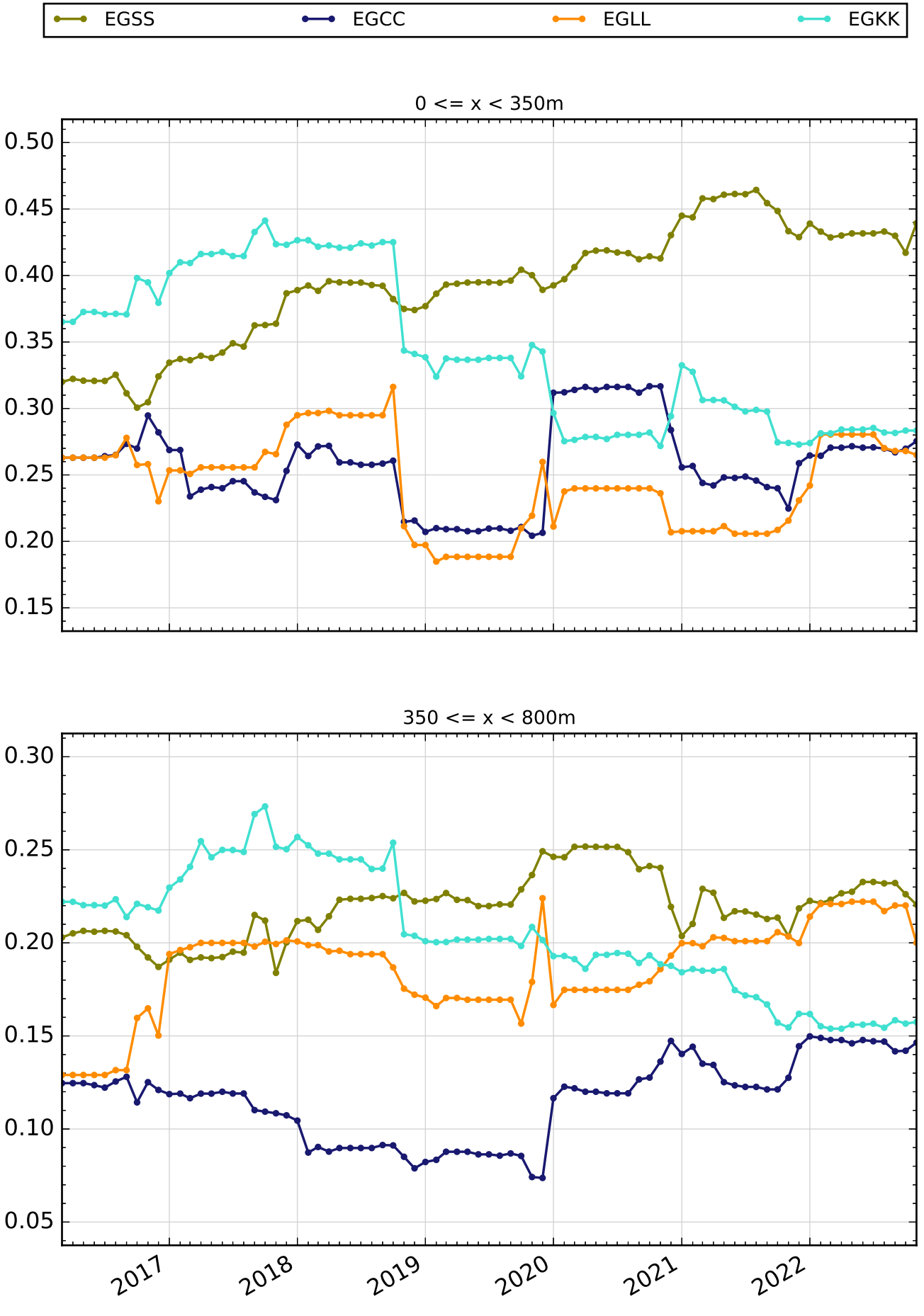
Airports used in verification

ICAO	Name	Duration	Region
EGCC	Manchester	30 hr	South
EGKK	London Gatwick	30 hr	South
EGLL	London Heathrow	30 hr	South
EGSS	London Stansted	30 hr	South

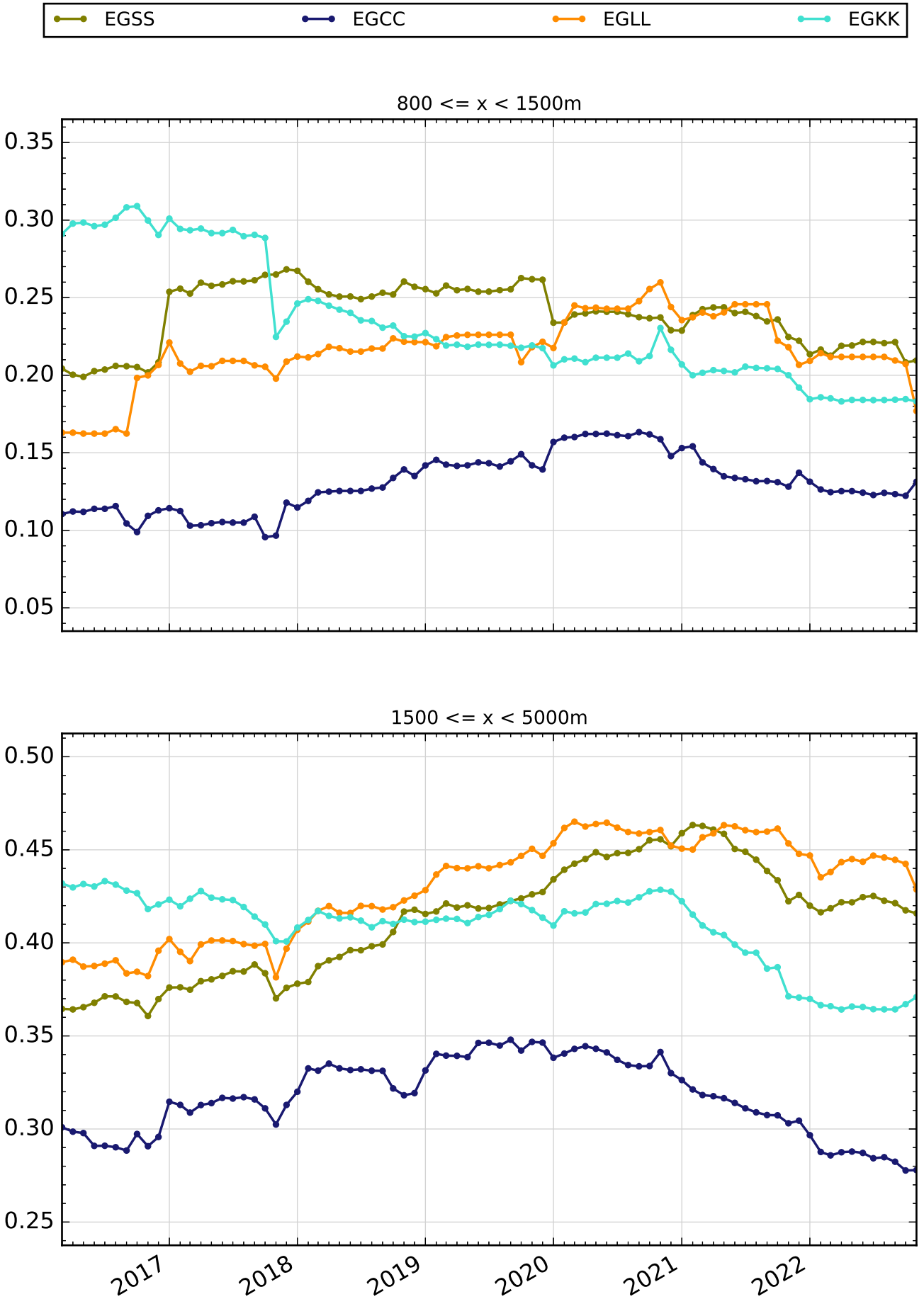
Visibility
Gerrity Skill Score
Rolling 36-month calculation



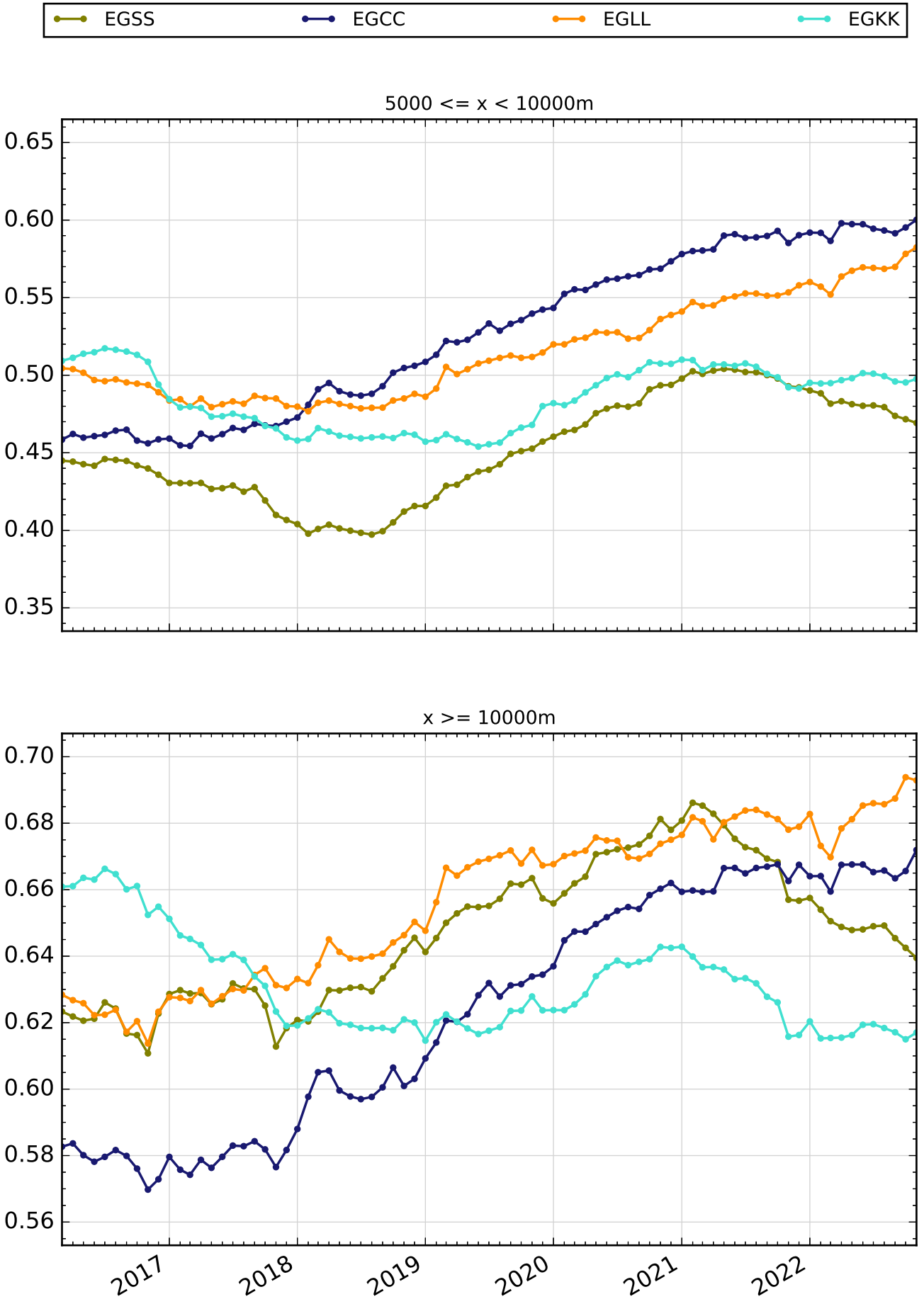
Visibility
Peirce Skill Score
Rolling 36-month calculation



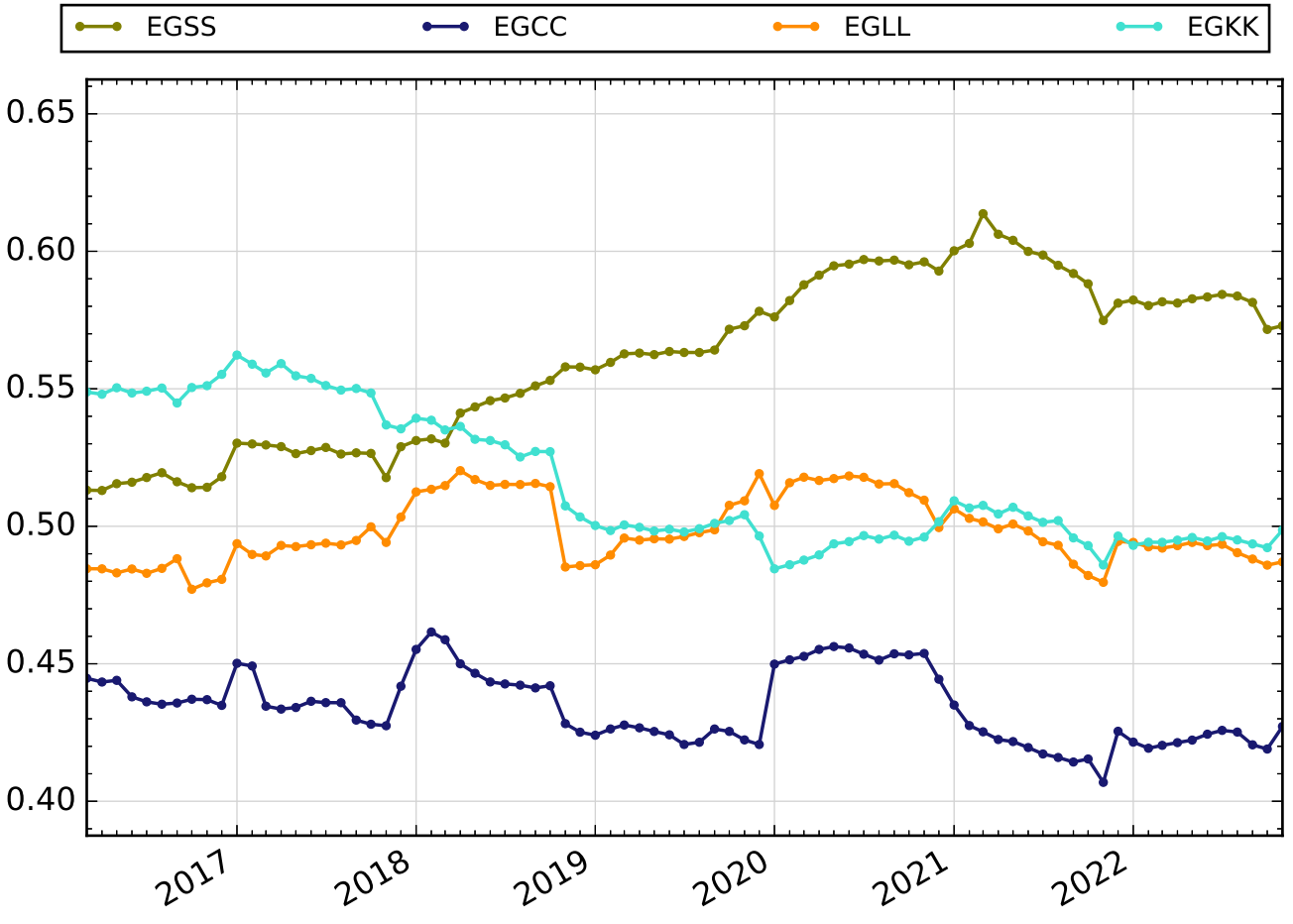
Visibility
Peirce Skill Score
Rolling 36-month calculation



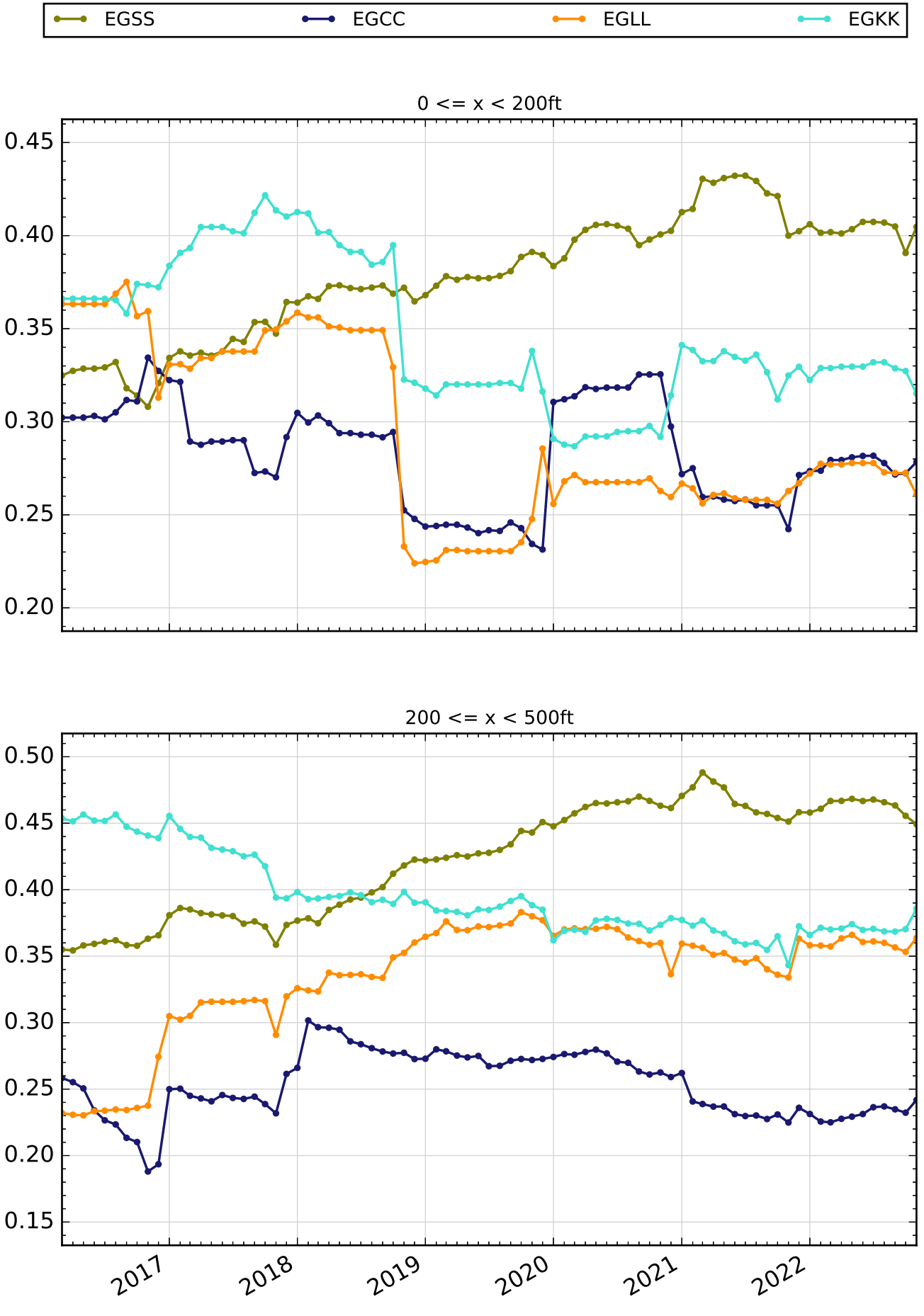
Visibility
Peirce Skill Score
Rolling 36-month calculation



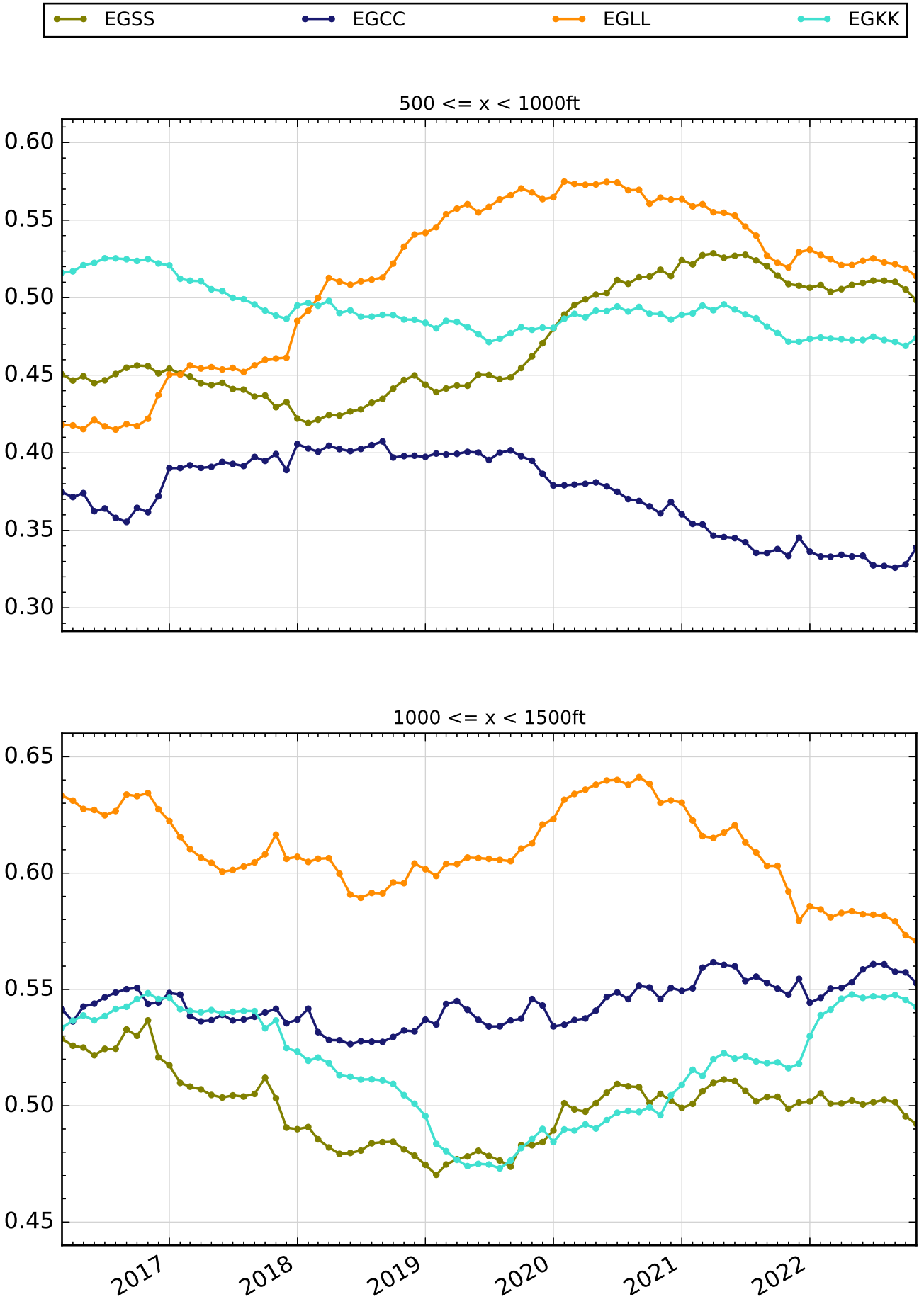
Cloud Base
Gerrity Skill Score
Rolling 36-month calculation



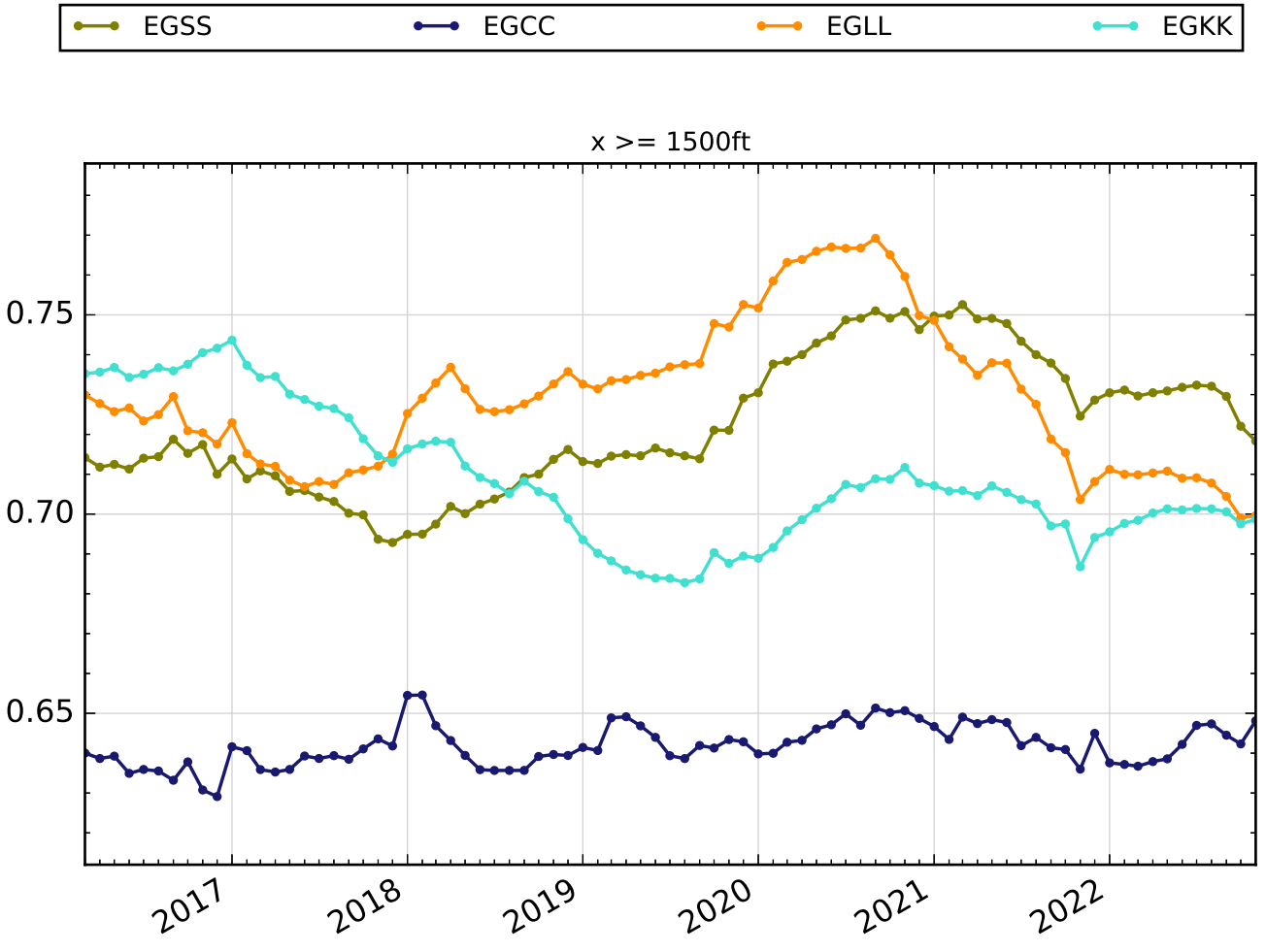
Cloud Base
Peirce Skill Score
Rolling 36-month calculation



Cloud Base
Peirce Skill Score
Rolling 36-month calculation



Cloud Base
Peirce Skill Score
Rolling 36-month calculation



Generated December 2022 for November 2022 results
© Crown copyright 2022

Met Office
FitzRoy Road, Exeter
Devon EX1 3PB
United Kingdom

Tel: 0370 900 0100
Fax: 0370 900 5050
enquiries@metoffice.gov.uk
www.metoffice.gov.uk