Table 1: QSA Uniform improves WIS by increasing interval widths.

method	interval score	dispersion	underprediction	overprediction
original	65.74	12.00	5.83	47.91
qsa_uniform	60.00	26.84	3.44	29.73
qsa_flexible_symmetric	60.92	33.22	2.49	25.21
qsa_flexible	60.47	25.31	9.31	25.84

0.1 Results

As for the CQR method, we investigate how well QSA performs for post-processing Covid-19 forecasts. We mainly focus on the UK Covid-19 Forecasting Challenge data set and only mention `qsa_uniform` results in the European Forecast Hub data due to computational restrictions.

0.2 Aggregate

We begin by examining the results of the `qsa_uniform` method and taking a high level view. Table 1 presents the performance on the validation set, aggregated over all *models*, *target types*, *horizons* and *quantiles*. `qsa_uniform` clearly improves the Weighted Interval Score as it drops by -8.7285906 percent. As expected post-processing makes the prediction intervals larger as the dispersion increases by a factor of 2.236078.

The increased intervals cover more observations and thereby reduce the under- and overprediction by -40.9820534 and -37.9538049. Interestingly while both decreases are similar in terms of relative performance increases, there absolute effects on the interval score differ substantially. The underprediction reduction decreases the WIS by -2.3886835 which amounts to a relative decrease of merely -3.6335836 percent, while the overprediction drops by -18.1834865 which in relative terms are -27.6600975 percent. The main driver behind the increasing in the intervals, is that they do not reach low enough. Thus by increasing the intervals and achieving better coverage of smaller observations, while at the same time sacrificing interval sharpness, `qsa_uniform` improves the WIS.

This finding, of the post processing methods increasing intervals, confirms the hypothesis that humans tend to be too confident in their own forecasts leading to narrow prediction intervals.

Figure Figure 1 shows the WIS changes of `qsa_uniform` for each *horizon* and *quantile* combination, aggregated by *models* and *target types*. `qsa_uniform` is beneficial for extreme quantiles at large horizons. however it also substantially overfits extreme quantiles at the horizon of 1. interestingly, near to no changes can be observed for the smaller prediction intervals lying within the 0.25 and 0.75 quantiles.

...

Figure Figure 1 revealed no significant adjustments for the inner confidence intervals. Due to the restriction of identical quantile spread adjustments for all quantiles, inherent to `qsa_uniform`, the optimization cannot differ in its post-processing of the various intervals. It could be the case that smaller intervals might need different adjustments than larger ones. This can be the case if humans have difficulty of intuitively grasping the concept of confidence intervals. In order to investigate this question, we examines the `qsa_flexible_symmetric` method. It allows the QSA adjustments to vary between intervals. Its only restriction is for adjustments to be symmetric, hence identical for each quantile pair, being the lower and upper bounds of symmetric intervals.

Table Table 1 presents the aggregated performance of `qsa_flexible_symmetric` on the validation set. The WIS remains lower in comparison to the original data, however it does lie above the v `qsa_uniform` by -7.3313515 percent. in the aggregate `qsa_flexible_symmetric` seems to overfit compared to the much more restrictive `qsa_uniform`. Further evidence of overfitting are that the dispersion increases even further with a factor of 2.7679146, and that the underprediction as well as overprediction drop even lower with -3.3338902 and -22.7022787 percent changes.

In Figure **??** the WIS changes of `qsa_flexible_symmetric` for each *horizon* and *quantile* pair show how this

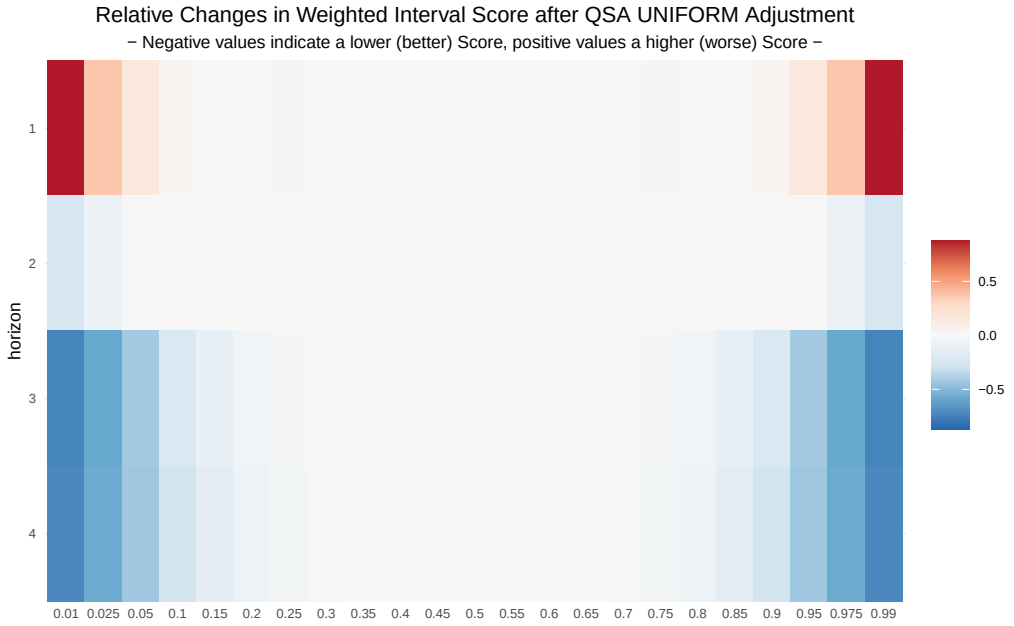


Figure 1: QSA Uniform beneficial for extreme Quantiles at large Horizons.

more flexible method adjusted the different intervals. Suprisingly we see no changes in the inner quantiles between the 0.3and 0.7 quantiles. Apparently the intervals with coverages equal or smaller than 50 percent where already quite optimal in the original human forecasts. Furthermore, the gains for the larger intervals remain similar, which suggests that the restriction to adjust all intervals with the same quantile spread factor, did not pose an issue for the Uk data set. In contrast, we rather observe an issue in the third horizon where more extreme quantile gains drop and extreme quantiles at lower horizon are overfitted by `qsa_flexible_symmetric` as the adjustments are worse than originally.

`qsa_uniform` and `qsa_flexible_symmetric` are both bound to symmetrically adjust upper and lower bounds of the prediction intervals. This is sensible for adjusting models who's residuals follow a symmetric distribution. If model residuals are however skewed, and thus interval coverage lacks more heavily on one side, symmetric adjustments lead to sub-optimal results. This happens because the model is confronted with a trade off where it adjusts one side to little and the other side to much. In the case where the post-processing increases intervals it is bound by the dispersion penalty that is heavier, since for each step it takes at reducing undercoverage, e.g. underprediction or overprediction, on one end, it increases dispersion two fold as intervals are also increased in the other end. In the case of decreasing intervals, it is bound by a lack of coverage as for each step it decreases unnecessary large intervals on one side, it also decreases the interval on the other side leading to uncovered observations. Thus, in both cases where post-processing is warranted, but the model residuals are non-symmetrical, symmetric methods lead to sub-optimal adjustments on both sides of the interval. As the Covid-19 infection and death data is inherently non-symmetrically distributed, due to the observations being bounded between $[0, Inf]$ and them resulting from exponential growth, we expect model residuals to be skewed towards higher values. Therefor, we examine how the non-symmetric post-processing method `qsa_flexible` adjusts the forecasts and how it preforms in contrast to `qsa_uniform` and `qsa_flexible_symmetric`.

Table 1 presents the aggregated performance of `qsa_flexible` on the validation set. The WIS is a clear improvement in comparison to the original data and lies in between the `qsa_uniform` and `qsa_flexible_symmetric`. Thus, it preforms slightly better than the `qsa_uniform` and slightly worse than the `qsa_flexible_symmetric` methods. Our main interest however lies in how intervals are adjusted, thus in the dispersion, underprediction and overprediction. The dispersion increases after post-processing, however to a lesser degree than for the other methods. The underprediction, most notably and in contrast to the symmetric approaches, substantially increases by 59.7643849 percent, while still remaning the lowest of the

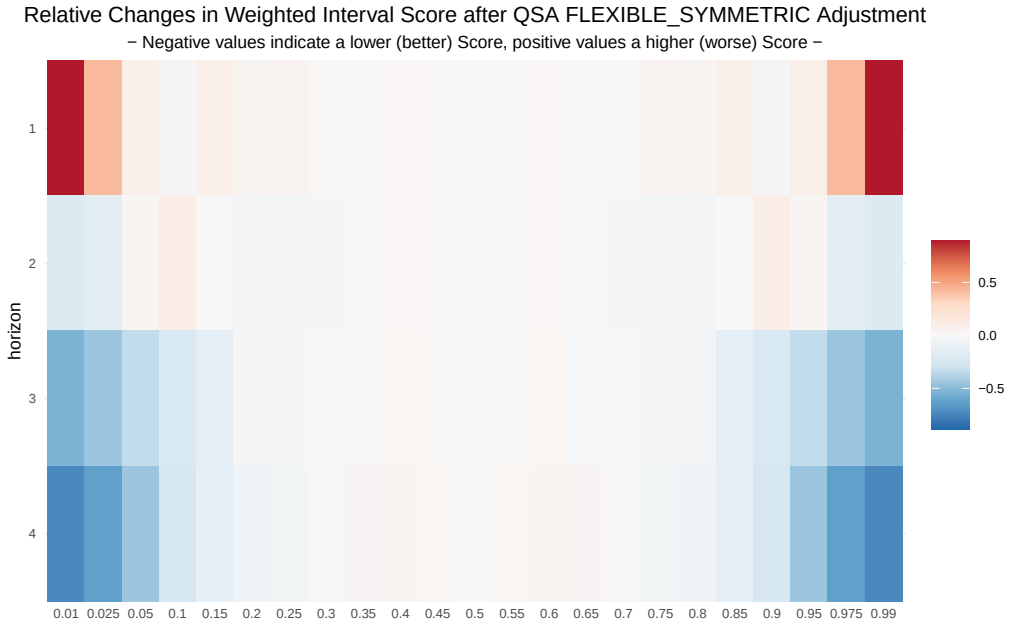


Figure 2: QSA Flexible Symmetric overfits Quantiles at short Horizons.

three WIS components. The overprediction behaves similarly to the `qsa_flexible_symmetric` method and decreases strongly by -46.0561394 percent. Due to the unsymmetrical nature of the misscoverage, in the aggregate, `qsa_flexible` moves the intervals downward, by heavily decreasing the lower quantiles in order to reduce overprediction and slightly decreasing the upper quantiles as the lost coverage is more than compensated by a reduction in dispersion. Surprisingly, due to the nature of exponential growth we would have expected human forecasters to underestimate trends, however for the UK Data, we observe an overconfidence in increasing cases and the death tool.

0.3 Across Models

?? shows the results of QSA post-processing for all three methods by model. These more granular results reveal a pattern not visible in the aggregated results. the `qsa_flexible` method preforms significantly worse in the `EuroCOVIDhub-baseline` model by increasing the WIS by 41.4127866 %. It overfits the training set as becomes evident by figure Figure 3. The figure shows that the original prediction intervals are below the actual values in the training period, then adjust to overshoot the actual values and level out during the validation period. As `qsa_flexible` is able to adjust the intervals in a non-symmetrical manner, it learns to push both interval bounds upward during the training set. As this pattern of underprediction changes in the validation set and the QSA metrics equally weigh all observations, `qsa_flexible` takes some observations to adjust properly and overpredicts the observations in the meantime. In contrast, `qsa_uniform` and `qsa_flexible_symmetric`, although the also don't improve the WIS in the validation set, overfit much less due to there restraint of making symmetric adjustments.

For the `EuroCOVIDhub-ensemble` model we observe that the `qsa_uniform` method has the best performance and could reduce the WIS by -2.331217. It seems that a simple, quite restrictive uniform adjustment across all quantiles provided the largest benefit. Adding additional flexibility among intervals with `qsa_flexible_symmetric` actually reduced the gains by about half and the further flexibility of `qsa_flexible` with non-symmetric adjustments even lead to a slightly worse prediction. These results are quite encouraging as the `EuroCOVIDhub-ensemble` model represents an ensemble of modelling approaches by professional forecasting models and thus isn't burdened by human overconfidence. For the human forecasting models, namely `epiforecasts-EpiExpert`, `epiforecasts-EpiExpert_Rt`, `epiforecasts-EpiExpert_direct` and `seabbs`, we observe that all QSA methods significantly improve the WIS. Furthermore for each model, there is at least one method that can reduce the Score by over 10%. Regarding the last three models we even see a

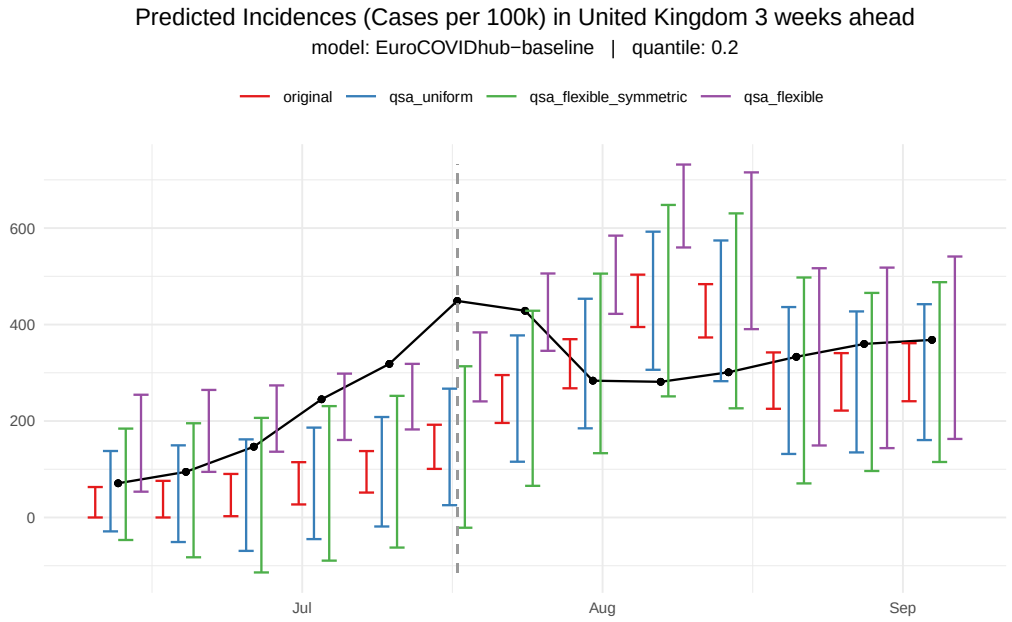


Figure 3: QSA Flexible overfits as it intervals are to low in the training and to high in the validation set.

similar pattern among post processing method performances: `qsa_flexible` reduces the WIS most, followed by `qsa_uniform` and `qsa_flexible_symmetric`. For the first method, this ranking is reversed, however, the scores only vary slightly. An inspection of the WIS components provides further interesting observations: `qsa_flexible` consistently reduces overprediction the most, is the only method that increases underprediction and has the lowest increase in dispersion. These observations are the result of the non-symmetric adjustments which allow `qsa_flexible` to reduce the lower bound without having to increase the upper one. For the optimization this has two effects: For one, it can decrease the lower bound much stronger as the cost of doing so, in terms of dispersion, are cut in half. Second, it can now freely adjust the upper quantile reduce its value until the increase in underprediction is balanced out with the reduction in dispersion.

0.4 Across Target Types

Comparing the QSA methods across target types reveals notable differences. Figure Figure 4 shows the relative changes in WIS after applying `qsa_uniform`, `qsa_flexible_symmetric` and `qsa_flexible` to the original data broken down by the `target_type`. In the aggregate all three methods improve the score for both target types within a similar range. `qsa_flexible_symmetric` preforms best for `Deaths` and `qsa_uniform` for `Cases`.

If we split models into human and model forecasts the results change as is depicted in Figure Figure 5. Human forecasts primarily benefit from post processing for the `target_type` `Cases`, while model forecasts are only improved in their `Deaths` predictions. For both major improvements, as discussed regarding the human forcasted models, `qsa_flexible` reduces the WIS most, followed by `qsa_uniform` and `qsa_flexible_symmetric`. in terms of overfitting, we observe that `qsa_flexible` is the only model that increases the score most. these results once more illustrate that `qsa_flexible` is a riskier model, as it can lead to higher gains or losses due to its potential to more closely fit the training data.

0.5 Across Horizons

Breaking down the results by the forecasting `horizon` also reveals notable patterns. Figure Section 0.5 plots the WIS changes across horizons for all three methods. Across all QSA methods the improvements to the WIS increase with the `horizon`. The gains are primarily visible for the three and four week ahead predictions. In Contrast the increases in score and overfitting are primarily located at a horizon of one.

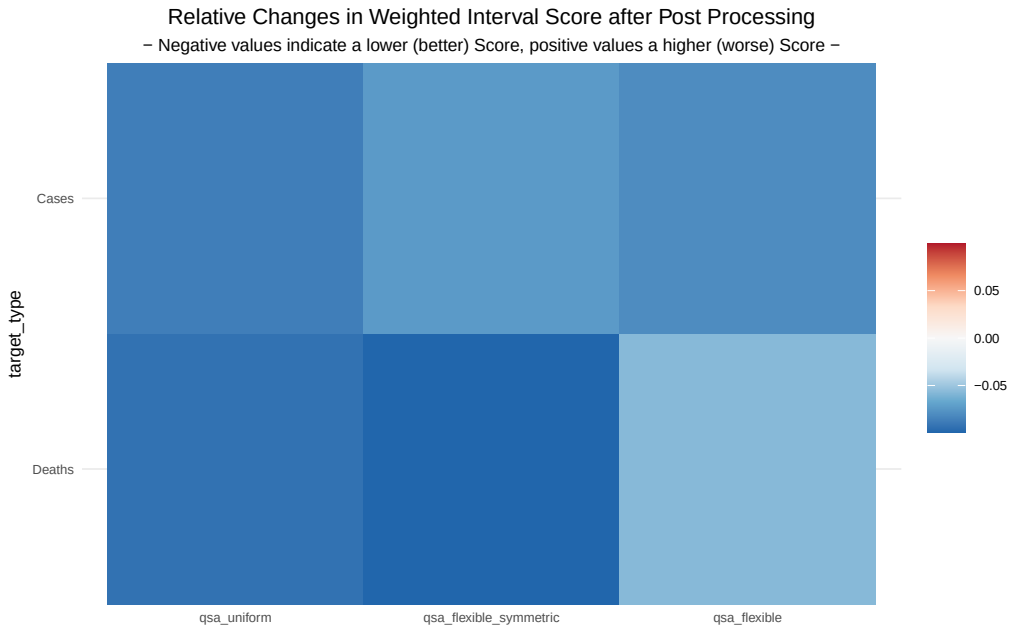
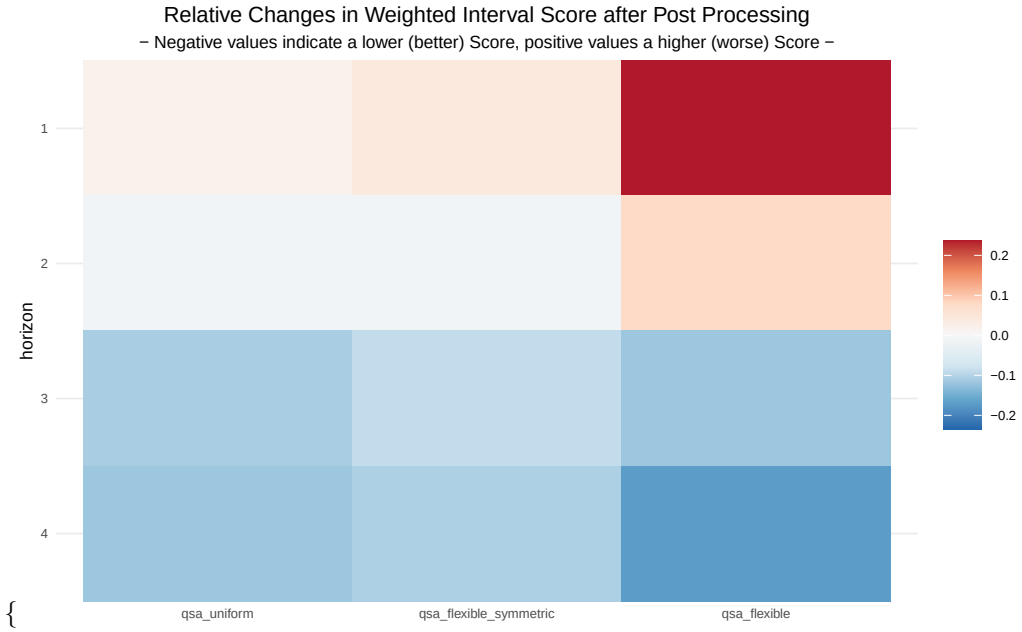


Figure 4: Across both Target Types all QSA mrthods improve the WIS.

\begin{figure}



{

}

\caption{QSA methods improve forecasts more for larger horizons. For smaller horizons they tend to voerfit, this is especially the case with qsa_flexible.} \end{figure}

Again, a split of the post processed models into human and model forecasts reveals differences as depicted in Figure **??**. We observe that the aggregate gains soley stem from the human forecast models and that the losses in the WIS are primarily from the model forecasts. Here the method performances also vary more. The largest gains and losses are reported for `qsa_flexible`, while `qsa_uniform` and `qsa_flexible_symmetric` also report improvements but overfit much less.

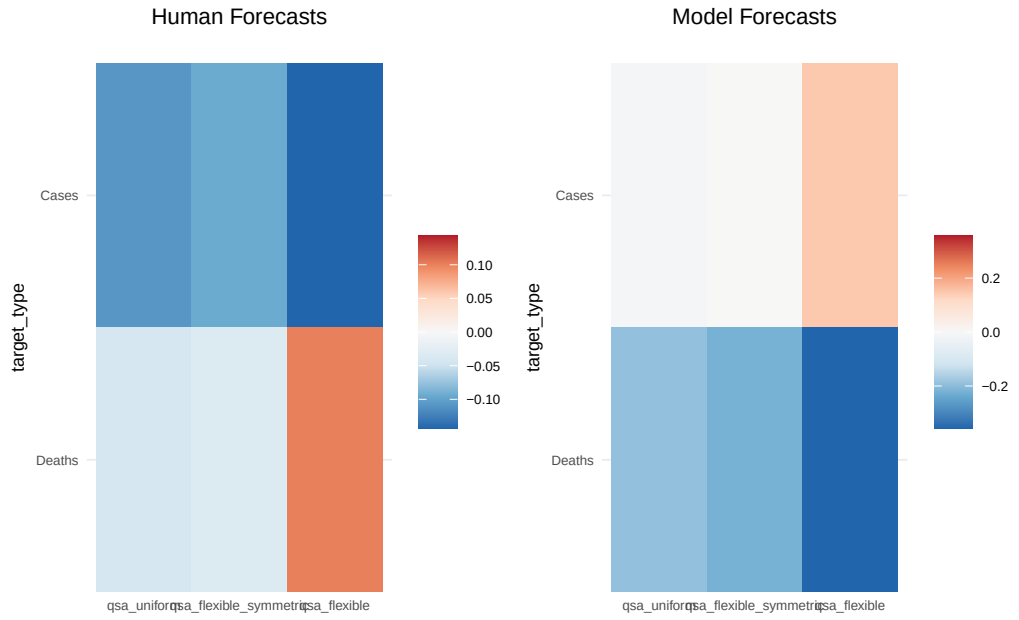
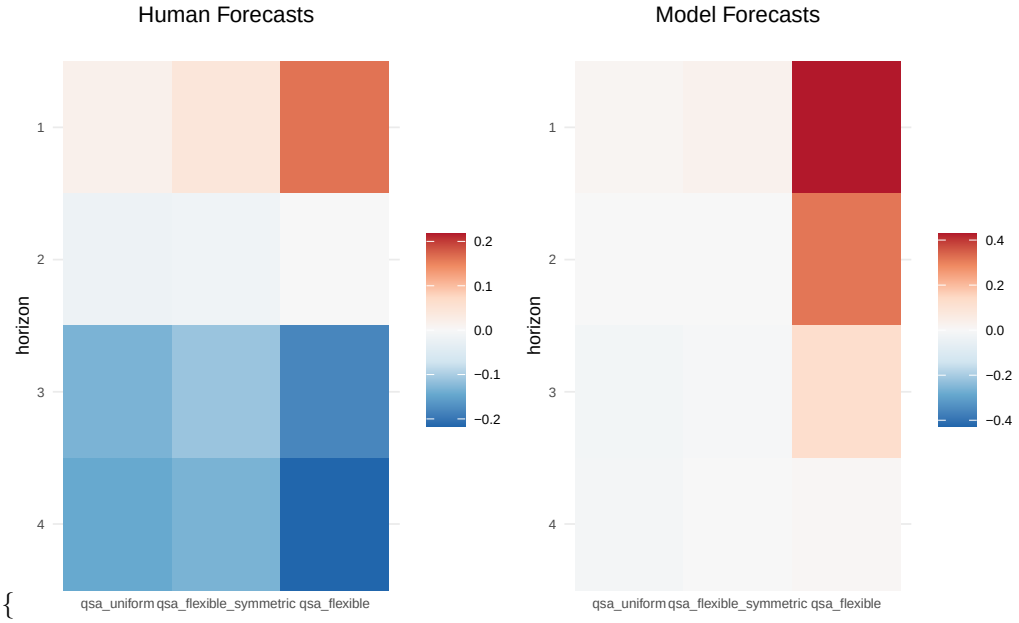


Figure 5: Forecasting improvements differ across target types for different model groups. Human forecasts are primarily improved for cases, while model forecast improvments are only found for deaths.

\begin{figure}



{

}

\caption{Forecasting improvements differ across horizons for different model groups. Human forecasts are primarily improved for horizons larger than 2, while model forecast are not improved at all and are overfitted with qsa_flexible.} \end{figure}

Additionally breaking down the results by target types as well, reveals the patterns in Figures **??** and **??** of the appendix. They show gains for death and losses for case predictions across the board for model forecasts. Again, these tendencies are strongest for the `qsa_flexible` flavor of QSA. Human predictions are primarily improved for cases and the horizons of 3 and 4 weeks ahead. For deaths `qsa_flexible` worsens the score,

especially for shorter horizons while `qsa_uniform` and `qsa_flexible_symmetric` slightly improve the scores across all horizons.

Overall there is a tendency that forecast improvements increase and and the risk of overfitting simultaneously drops with increasing forecasting horizons.

0.6 Across Quantiles

WIS improvements also vary across the different quantiles and intervals. Figure Figure 6 depicts the WIS across quantiles for the three QSA flavors. the improvements are larger for more extreme quantiles.

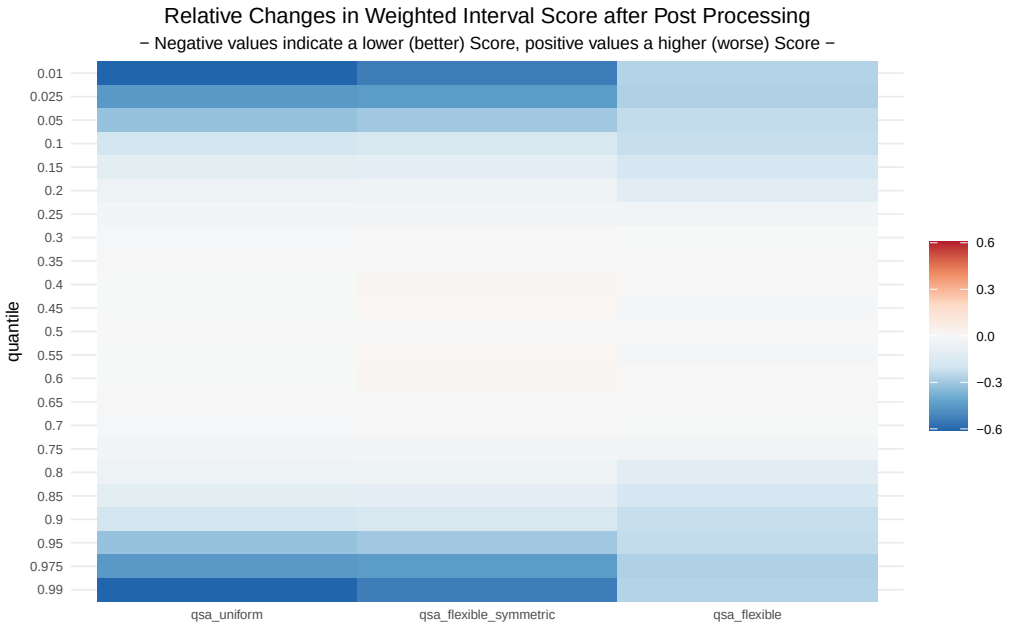


Figure 6: QSA improves forecasters larger for more extreme qunatiles and thus larger intervals.

As we have shown that aggregation across the `target_type`and `model` dimensions do not show the full picture, we also show the quantile improvements for the human forecasts of cases as well as the model forecasts of deaths in Figure Figure 7. For human forecasts of cases the patterns remain similar to the aggregate. For model forecasts of deaths however, we observe larger improvements and that `qsa_flexible` is useful for small intervals. This results from the death prediction intervals beenfitting from non-symmetric adjustments. Furthermore model forecasts of deaths also seem to be one of the rare cases where `qsa_flexible_symmetric` outpreforms `qsa_uniform`.

Further examination of the results, in particular subsetting the above results to the `horizon`of three and four revealed an exception to the accordion look of the quantile graphs. For large forecasting horizons and model forecasts of deaths, we observe worse WIS after the adjustments. These result from the high cost of not covering an observation at extreme quantiles. Figure Figure 8 exemplifies this where all QSA methods substantially reduces the interval sizes in order to reduce dispersion, this then result in undercoverage of the last week of august. Thus the QSA adjustments, especially for few data points ,as the 13 weeks of the UK data, can underestimate uncertainty at extreme quantiles. This risk increases with the flexibility of the QSA flavor.

```
## # A tibble: 2 x 4
##   quantile qsa_flexible qsa_flexible_symmetric qsa_uniform
##      <dbl>        <dbl>                  <dbl>       <dbl>
## 1     0.01         3.65                  0.958        3.99
## 2     0.99         3.65                  0.958        3.99
```

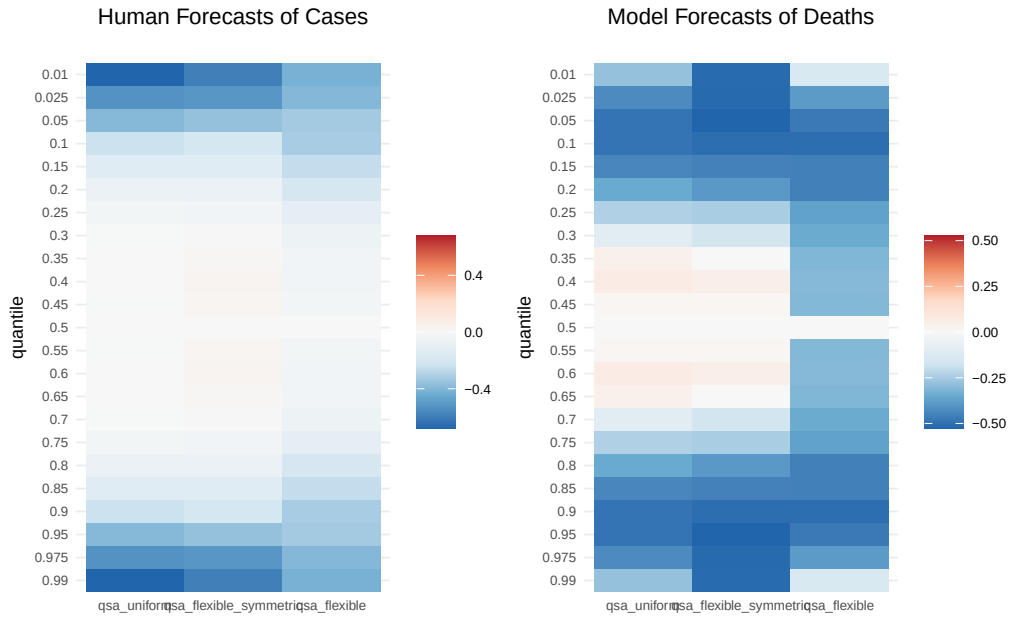


Figure 7: QSA Flexible overfits as it intervals are to low in the training and to high in the validation set.

```
## # A tibble: 2 x 2
##   model                     n
##   <chr>                 <int>
## 1 EuroCOVIDhub-baseline  4784
## 2 EuroCOVIDhub-ensemble  4784
```

0.7 Conclusion

- qsa uniform overall most consistent model
- symmetric flexible didnt bring notable gains
- qsa flexible works better or worse depending on the case
- human forecaster models more improved ofr cases and forecaster models rather for deaths
- regarding horizons and qunatile in general the further out the better
- talk about improvements and next steps

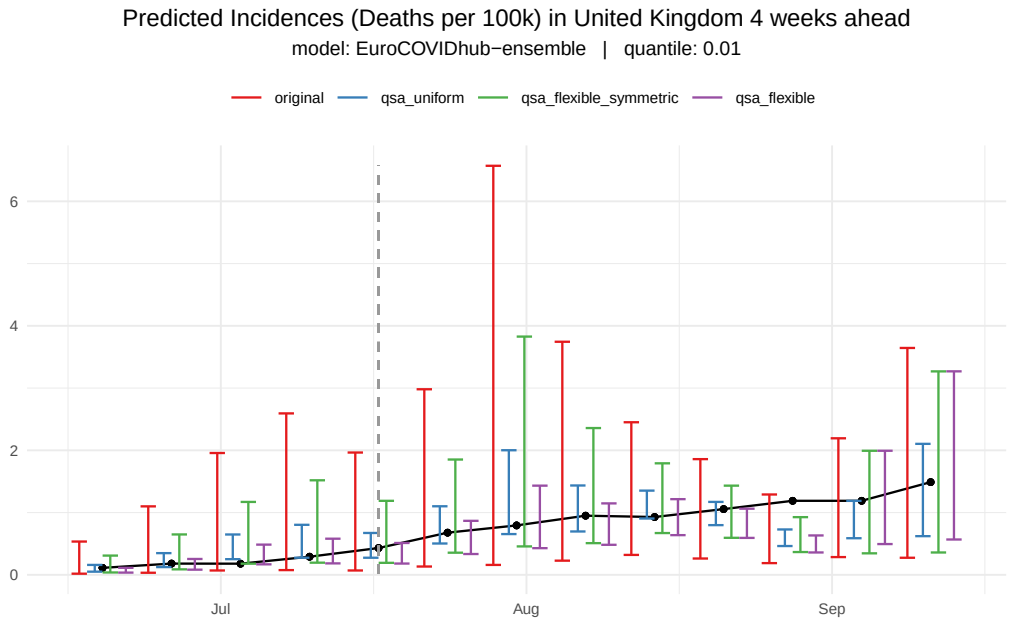


Figure 8: QSA can underestimate uncertainty for extreme qunatiles and few data points to learn from.