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 focus on the UK Covid-19 Forecasting Challenge data set due to computational restrictions.

0.1.1 Aggregate

We begin by examining a high-level overview of the results. Table 1 presents the performance of all three QSA flavors on the validation set, aggregated over all *models*, *target types*, *horizons* and *quantiles*.

Starting with `qsa_uniform`, we observe clear improvements in the WIS as it drops by -8.73%. As expected, this first update increases the overall dispersion and hence the weighted prediction intervals by a factor of 2.24. The wider intervals therefore cover more observations which reduces the under- and overprediction by -40.98% and -37.95%.

Interestingly, while both decreases are similar in relative terms, their *absolute* effects differ substantially. The underprediction reduction decreases the WIS by -2.39 which amounts to a relative WIS decrease of merely -3.63 %, while the overprediction drops by -18.18 which is equal to a WIS drop by -27.66 %.

We can thus conclude that the main driver behind the increase of the intervals is their *over*coverage. In other words: the intervals do not reach low enough. Overall, by increasing the intervals and achieving better coverage of smaller observations, while at the same time sacrificing interval sharpness, `qsa_uniform` improves the WIS.

Due to the restriction of identical quantile spread adjustments for all quantiles, the optimization cannot differ in its post-processing of the various intervals. We speculate 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, especially since we have seen that the adjustments of `qsa_uniform` are quite substantial. This line of reasoning is our motivation behind the `qsa_flexible_symmetric` method. It allows the QSA adjustments to vary between intervals. Its only restriction is that the updates with regards to the lower and upper bounds must be *symmetric.*

Table 1 also presents the aggregated performance of `qsa_flexible_symmetric`. It reports that the WIS remains lower in comparison to the original data, however it lies above the value of `qsa_uniform` by -7.33%. In the aggregate `qsa_flexible_symmetric` seems to overfit compared to the much more restrictive `qsa_uniform`. Further evidence of overfitting is that the dispersion continues to increase with a factor of 2.77 and that the underprediction as well as overprediction drop even lower by -3.33% and -22.7%.

`qsa_uniform` and `qsa_flexible_symmetric` are both restricted to symmetrically adjust upper and lower bounds of the prediction intervals. This is sensible for adjusting models producing residuals that follow a symmetric distribution. If, in contrast, model residuals are *skewed*, and thus the 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 too little and the other side too much.

In the case where the post-processing *increases* intervals it is bound by the dispersion penalty that is heavier, since for each step that it takes to reduce undercoverage, e.g. underprediction or overprediction, on one end, it increases dispersion two fold as intervals are also increased on 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

Table 2: QSA Methods differ in performance across Models.

model	**method**	**wis**	**dis**	**under**	**over**
EuroCOVIDhub-baseline	uniform	1.41	151.79	-38.53	-33.79
EuroCOVIDhub-baseline	symmetric	2.26	240.03	-60.99	-53.30
EuroCOVIDhub-baseline	flexible	41.41	151.80	-85.95	116.68
EuroCOVIDhub-ensemble	uniform	-2.33	50.93	-26.07	-18.72
EuroCOVIDhub-ensemble	symmetric	-1.03	86.75	-35.61	-29.04
EuroCOVIDhub-ensemble	flexible	0.49	51.77	26.47	-26.16
epiforecasts-EpiExpert	uniform	-10.91	110.29	-49.31	-36.45
epiforecasts-EpiExpert	symmetric	-10.33	150.09	-67.84	-43.74
epiforecasts-EpiExpert	flexible	-10.12	119.72	118.81	-47.80
epiforecasts-EpiExpert_Rt	uniform	-8.74	120.80	-48.07	-36.26
epiforecasts-EpiExpert_Rt	symmetric	-7.59	184.67	-62.42	-48.76
epiforecasts-EpiExpert_Rt	flexible	-18.07	91.64	153.22	-60.41
epiforecasts-EpiExpert_direct	uniform	-10.43	122.49	-47.78	-38.19
epiforecasts-EpiExpert_direct	symmetric	-9.78	154.27	-50.21	-44.34
epiforecasts-EpiExpert_direct	flexible	-12.15	124.66	145.45	-50.99
seabbs	uniform	-12.91	199.50	-56.50	-49.32
seabbs	symmetric	-9.98	273.52	-77.04	-58.13
seabbs	flexible	-15.62	153.56	338.63	-63.88

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-symmetric, 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 since the observations are bounded between $[0, Inf]$ and result from exponential growth, we expect model residuals to be skewed towards higher values. Therefore, 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 between `qsa_uniform` and `qsa_flexible_symmetric`. Thus, it performs 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, i.e. in the dispersion, underprediction and overprediction.

The dispersion increases after post-processing, yet to a lesser degree than for the other methods. The underprediction, most notably and in contrast to the symmetric approaches, substantially increases by 59.76%, while still remaining the lowest of the three WIS components. The overprediction behaves similarly to the `qsa_flexible_symmetric` method and decreases strongly by -46.06%.

Due to the asymmetric nature of the miscoverage, `qsa_flexible` moves the intervals downwards in the aggregate 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. 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.

In the following subsections we increase the granularity of our analysis and examine the QSA flavor preformances across the dimensions of our data, namely the *models*, *target types*, *horizons* and *quantiles*.

0.1.2 Models

Table 2 displays the results of QSA post-processing for all three methods stratified by model. These more granular results reveal a pattern that is not visible in the aggregated results: The `qsa_flexible` method performs significantly worse for the `EuroCOVIDhub-baseline` model by increasing the WIS by 41.41%. It overfits the training set as indicated by Figure 1.

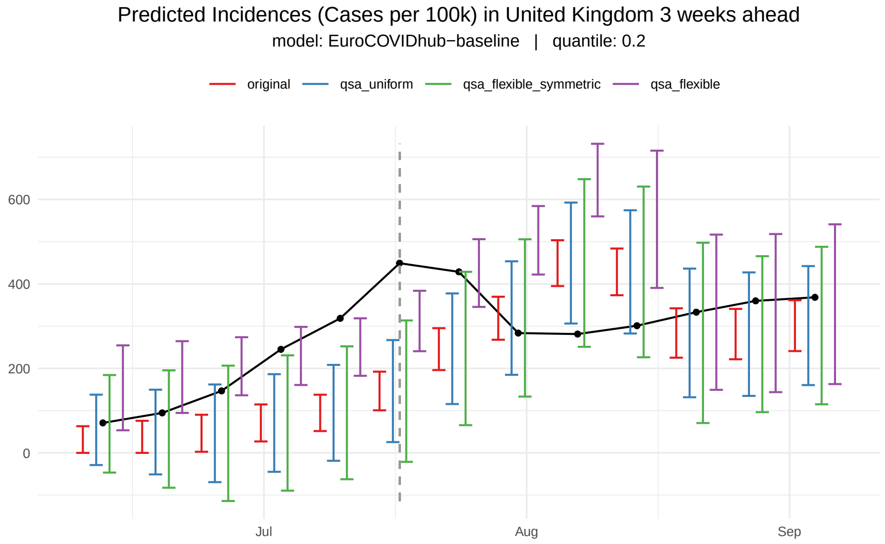

Figure 1: QSA Flexible overfits as it intervals are to low in the training and to high in the validation set.

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 upwards 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` overfit much less due to their constraint of symmetric adjustments, yet they similarly do *not* improve the WIS in the validation set.

For the `EuroCOVIDhub-ensemble` model we observe that the `qsa_uniform` method has the best performance and reduces the WIS by -2.33. It seems that a simple, quite restrictive uniform adjustment across all quantile levels provides the largest benefit. Adding additional flexibility among intervals with `qsa_flexible_symmetric` actually *reduces* the gains by about half and the further flexibility of `qsa_flexible` with non-symmetric adjustments even leads to a slightly worse prediction. These results are quite encouraging as the `EuroCOVIDhub-ensemble` model represents an ensemble of modeling approaches by professional forecasting models and is therefore not 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 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, although the scores vary only slightly. An inspection of the WIS components provides further insight: `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 counterpart. For the optimization this has two effects: First, it can decrease the lower bound much stronger since the cost in terms of dispersion is cut in half. Second, it can now freely adjust the upper quantile until the increase in underprediction is balanced out with the reduction in dispersion.

0.1.3 Target Types

Comparing the QSA methods across target types reveals notable differences. Figure 2 shows the relative changes in WIS after applying `qsa_uniform`, `qsa_flexible_symmetric` and `qsa_flexible` to the original data broken down by `target_type`. In the aggregate all three methods improve the score for both target types within a similar range. `qsa_flexible_symmetric` performs best for Covid-19 Deaths and `qsa_uniform` for Cases.

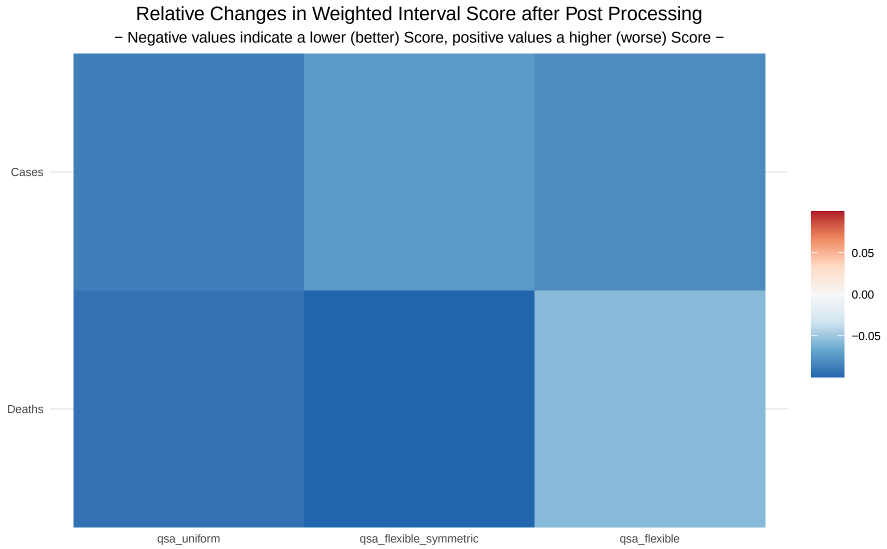

Figure 2: Across both Target Types all QSA methods improve the WIS.

If we split the forecasting models into human and non-human groups the results change as is shown by Figure 3. Human forecasts primarily benefit from post processing Covid-19 `Cases`, while model forecasts are only improved in their Deaths predictions.

As discussed for the human forecast 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 model that increases the score most. These results illustrate once again that `qsa_flexible` is a riskier model, as it can lead to higher gains or losses due to its potential to fit the training data too closely.

0.1.4 Horizons

Breaking down the results by the forecasting `horizon` also reveals notable patterns. Figure 4 plots the WIS changes across horizons for all three methods. Across all QSA methods the improvements of the WIS increase with the `horizon` level. 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.

Again, a split of the post processed models into human and model forecasts reveals differences as shown by Figure 5. We observe that the aggregate gains solely stem from the human forecasts 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 indicate improvements but overfit much less.

Additionally, breaking down the results by target types reveals the patterns in **??** and **??** in 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 Covid-19 Cases and forecast horizons of 3 and 4 weeks. For Covid-19 Deaths `qsa_flexible` worsens the

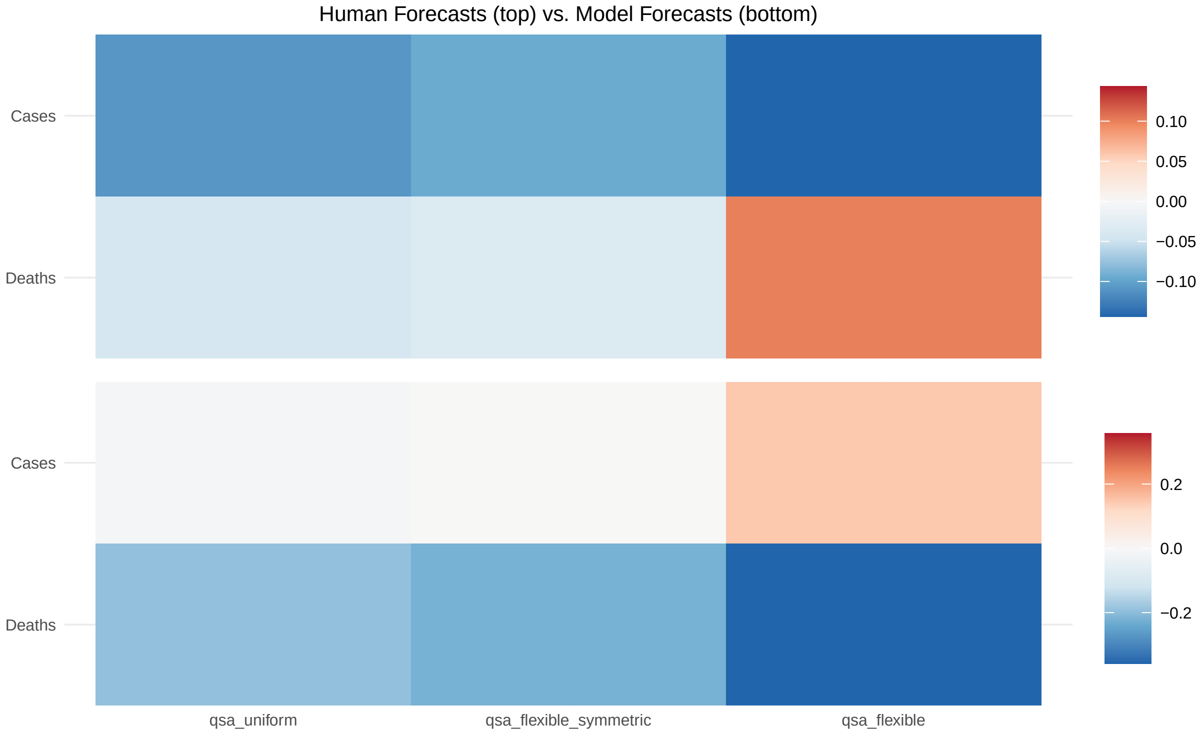

Figure 3: Forecasting improvements differ across target types for different model groups. Human forecasts are primarily improved for Covid-19 Cases, while model forecast improvments are only found for Covid-19 Deaths.

score for shorter horizons in particular, while `qsa_uniform` and `qsa_flexible_symmetric` slightly improve the scores across all horizons.

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

0.1.5 Quantiles

WIS improvements also vary across different quantile levels. As shown in Figure 6 the improvements are larger for more extreme quantiles. For smaller prediction intervals between the 0.25 and 0.75 quantiles almost no improvements can be observed for any of the three flavors.

This finding is particularly important regarding `qsa_flexible_symmetric`, as the main motivation behind it was to allow for individual adjustments of each interval pair. Thus, we would have expected `qsa_flexible_symmetric` to perform better, particularly for intervals which `qsa_uniform` could not improve due to its restrictive nature. Apparently, the intervals with coverage equal or smaller than 50% were 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.

Yet, aggregation across the `target_type`and `model` dimensions do not represent the full picture, such that we also show the quantile improvements for the human forecasts of Cases as well as the model forecasts of Deaths in Figure 7. For human forecasts of Covid-19 Cases the patterns remain similar to the aggregate. For model forecasts of Covid-19 Deaths, however, we observe larger improvements and detect that `qsa_flexible` is useful for small intervals. This suggests that Death prediction intervals benefit from non-symmetric adjustments. Furthermore, model forecasts of Deaths also seem to be one of the rare situations where `qsa_flexible_symmetric` outperforms `qsa_uniform`.

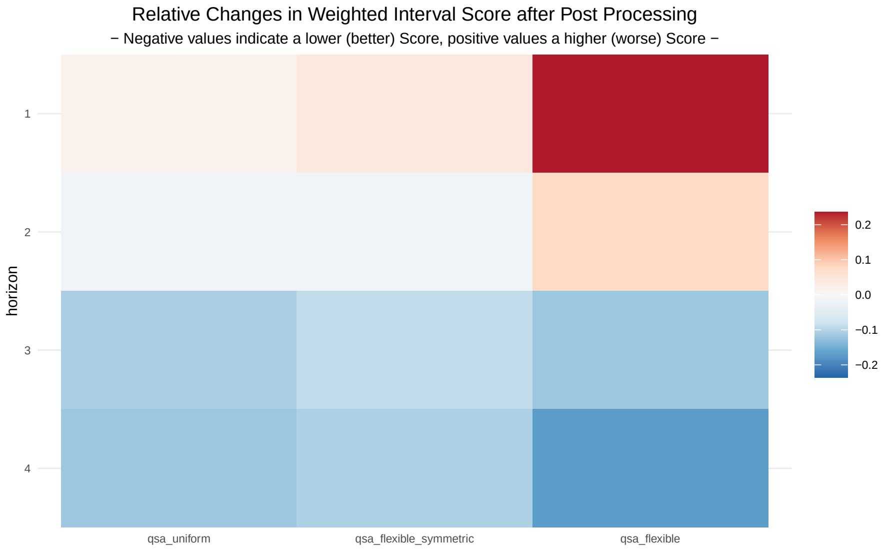

Figure 4: QSA methods improve forecasts more for larger horizons. For smaller horizons they tend to overfit, this is especially the case with QSA Flexible.

Subsetting the above results to a forecast horizon of three and four weeks reveals an exception to the look of the quantile graphs. For large forecast horizons and model forecasts of Deaths, we observe worse WIS after the adjustments. These stem from the high cost of not covering an observation at extreme quantiles. Figure 8 exemplifies this where all QSA methods substantially reduce the interval sizes in order to decrease dispersion, which then results in undercoverage of the last week of August.

Thus, the QSA adjustments can underestimate uncertainty at extreme quantiles, especially for short time series as the 13 weeks of the UK data. This risk increases with the flexibility of the QSA flavor.

0.1.6 Conclusion

Overall `qsa_uniform` is the QSA flavor that performs best for the UK data set. It produces notable improvements to the WIS in the validation set without overfitting the training data. The additional flexibility among interval adjustments that `qsa_flexible_symmetric` provides does not lead to significant gains.

Most surprisingly, `qsa_uniform` can not improve the WIS for smaller prediction intervals. It rather has the tendency to slightly overfit the data. The additional flexibility of non-symmetric interval adjustments offered by `qsa_flexible` have less clear effects. Overall, `qsa_flexible` and `qsa_flexible_symmetric` can not lower the WIS more than `qsa_uniform`, yet `qsa_flexible` outpreforms the other methods in the scenarios where post processing is most useful. `qsa_flexible` substantially overfits the data due to its non-symmetric adjustments as became evident for the `EuroCOVIDhub-baseline` model.

In general `qsa_uniform` is the more conservative choice, while `qsa_flexible` can be a better fit for data requiring large and varying adjustments across quantiles. With regards to the question when to use QSA, it performes best when forecasts underestimat uncertainty which was the case for larger horizons as well as more extreme quantiles. Furthermore, the method performance also depends on the types of forecasting models as well as the target type. Both taken separately, QSA worked best for human forecast models and Covid-19 Case predictions.

Observed together we did find that QSA performed best for the human forecasts of Cases and model forecasts of Deaths. The former observations are in line with the hypothesis that humans can not grasp uncertainty as well as models and that forecasting uncertainty is higher for Covid-19 Cases than for Deaths, since Deaths are strongly linked to *past* Cases.

Finally, these results leave much room for further investigations and improvements to the QSA method. First,

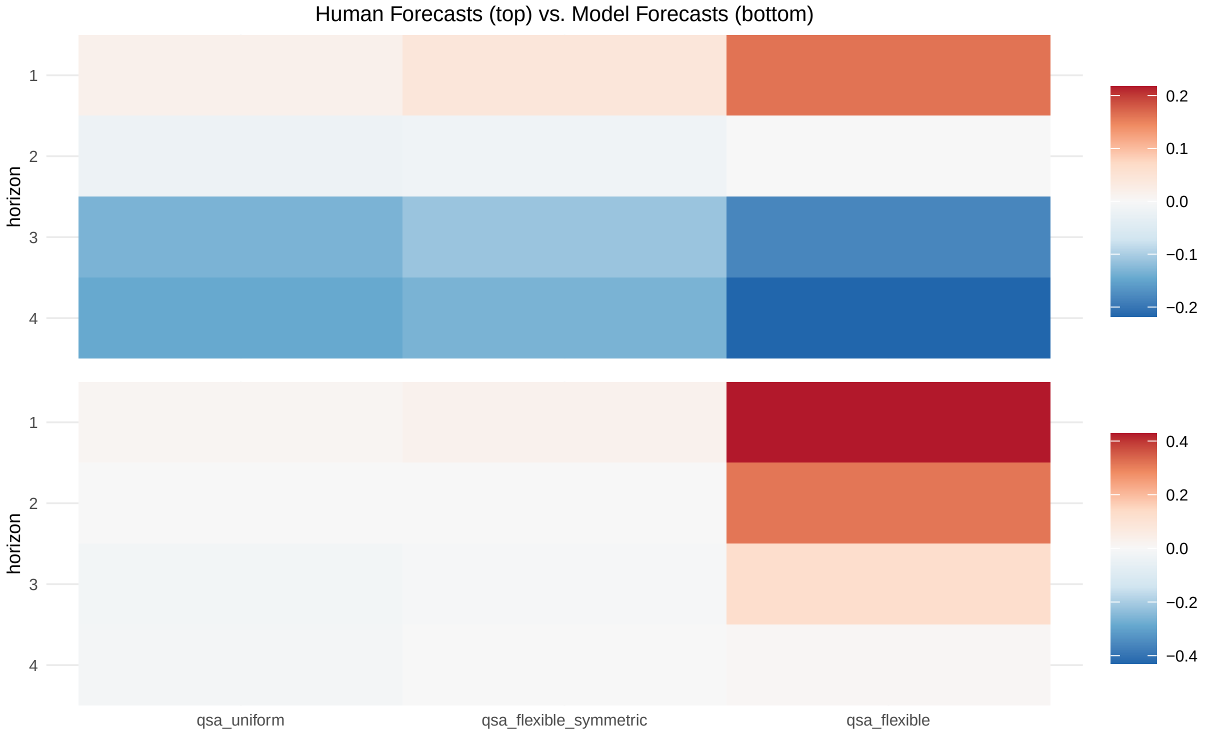

Figure 5: Forecasting improvements differ across horizons for different model groups. Human forecasts are primarily improved for horizons larger than 2, while model forecast are not improved at all and are overfitted with QSA Flexible.

it would be interesting to see whether the described results generalize to the European Forecast Hub data set. This would be particularly interesting as it contains longer time series and hence the models have more data to learn from, which could be an advantage for the more flexible methods. Second, a natural step would be to apply `qsa_flexible_symmetric` and `qsa_flexible` with *penalization*. This option requires keeping some observations in a separate test set as we would fit the penalty value to the validation set.

In regards to additional methods, one could imagine an asymmetric version of `qsa_uniform` with one adjustment for the quantiles below and one for those above the median. With this method one could investigate whether gains observed with `qsa_flexible` can be attributed solely to the asymmetry or are rather to the flexibility across interval levels.

In general, it would be desirable to allow users to set custom restrictions to the vector of QSA factors. Regarding penalization, it might also be interesting to add a penalization form that penalizes towards no adjustments, i.e. pulling all QSA factors closer to a value of one. Moreover, one could weigh the importance of observations in the optimization by their time point. Therefore we suggest an *exponential smoothing* approach that could assign larger weights to more recent observations which would allow the optimization of QSA to adapt faster to changes in the data. This modification would introduce a smoothing hyperparameter that, similar to the penalty approach, requires an additional learning step on the validation set.

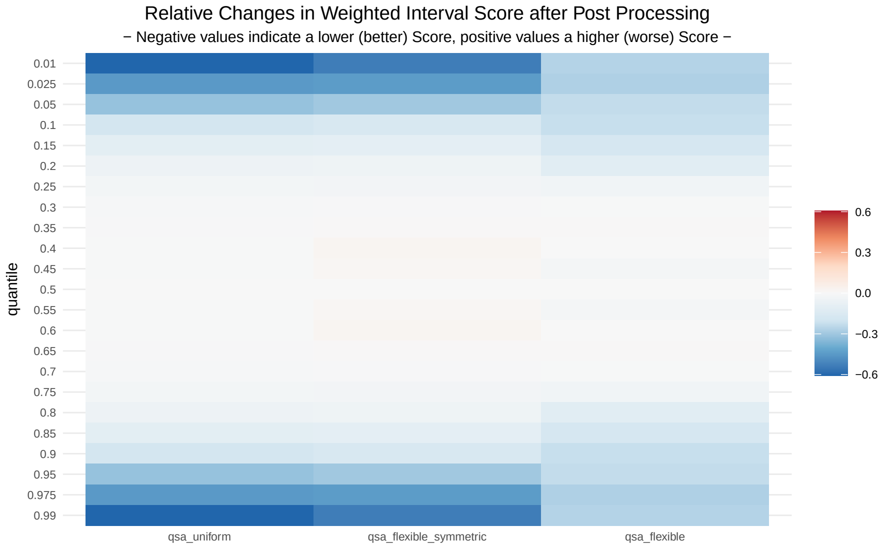

Figure 6: QSA improves forecasters more for more extreme quantiles and thus larger intervals.

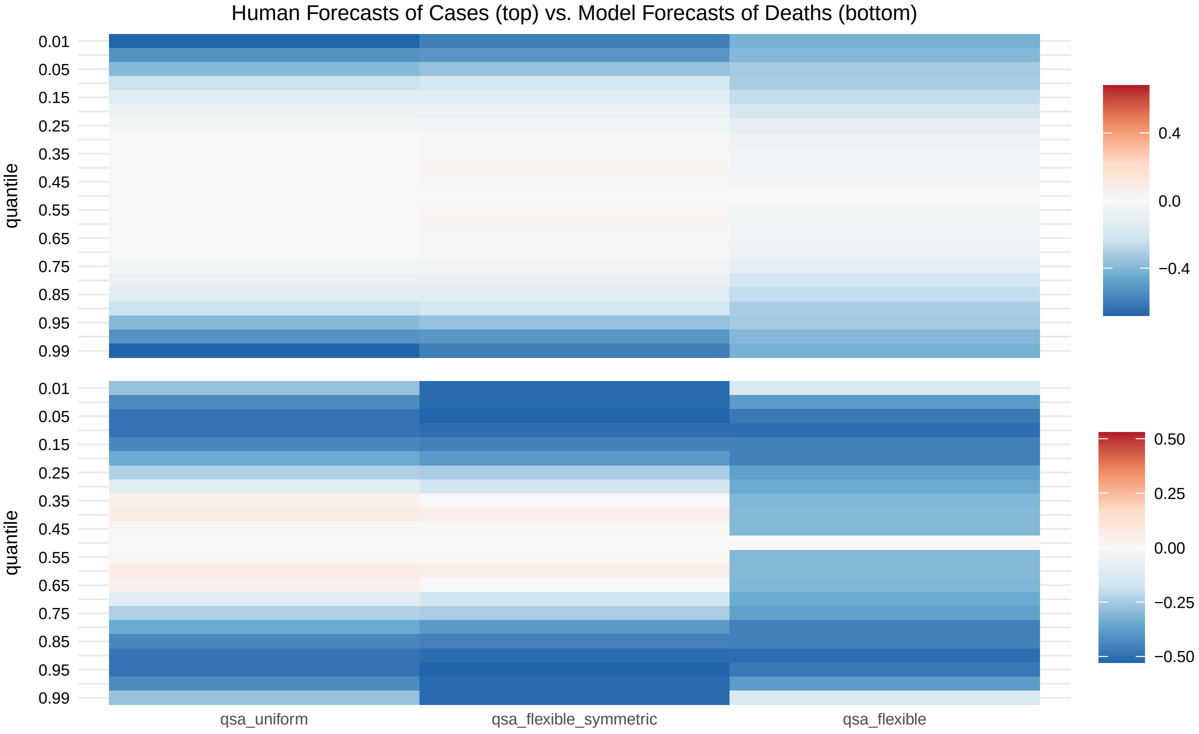

Figure 7: QSA Flexible overfits as it intervals are to low in the training and to high in the validation set.

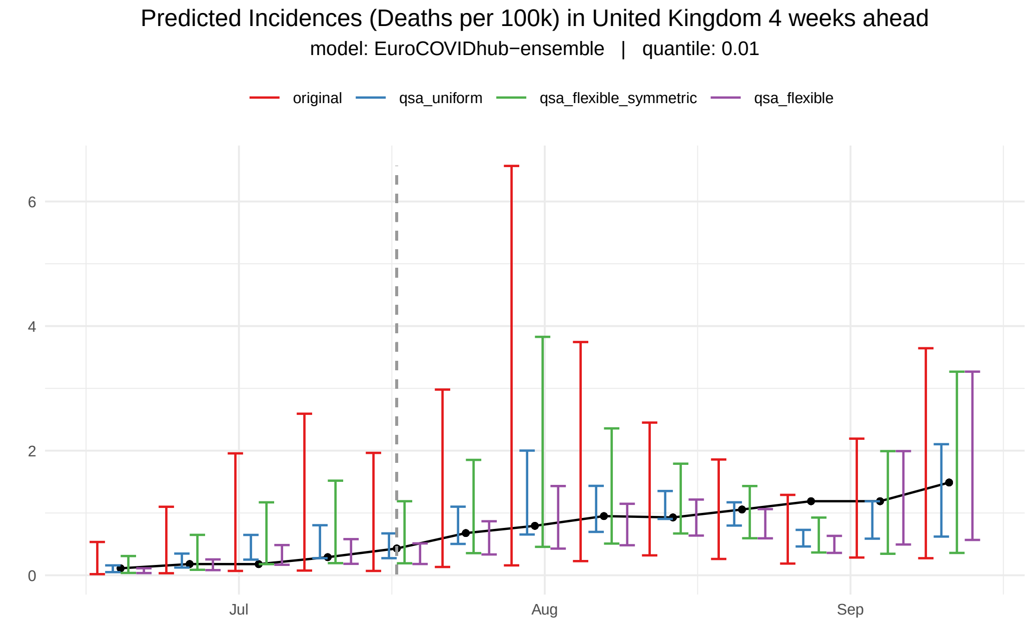

Figure 8: QSA can underestimate uncertainty for extreme quantiles and few data points to learn from.