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 clearl improvements in the WIS as it drops by -8.73 %. As expected, overall, the post-processing increases the dispersion and hence the weighted prediction intervals by a factor of 2.24. The wider intervals thereby cover more observations which reduces the under- and overprediction by -40.98 %. and -37.95 %. Interestingly, while both decreases are similar in relative terms, there 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 %. The main driver behind the increase of the intervals, thus is there overcoverage. 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, inherent to `qsa_uniform`, 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 for adjustments to be symmetric, hence identical for each quantile pair, being the lower and upper bounds of symmetric intervals.

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 does lie above the v `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 are that the dispersion increases even further with a factor of 2.77, and that the underprediction as well as overprediction drop even lower with -3.33 and -22.7 % changes.

`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 producing residuals that 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

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

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 59.76 %, while still remaning 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 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.

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 depicts 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 41.41%. It overfits the training set as becomes evident by Figure 1. 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.

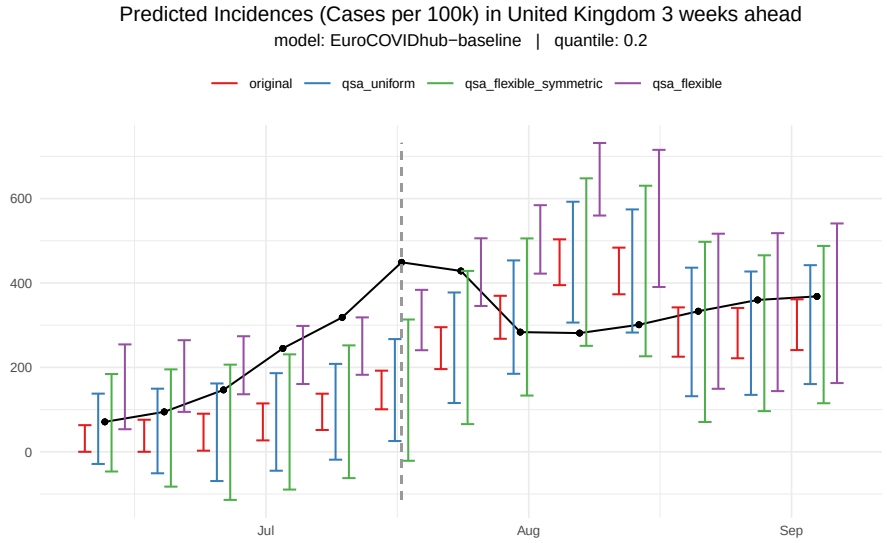

Figure 1: QSA Flexible overfits as it intervals are to low in the training and to high in the validation set.

For the `EuroCOVIDhub-ensemble` model we observe that the `qsa_uniform` method has the best performance and could reduce the WIS by -2.33. 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 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.

0.1.3 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 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`.

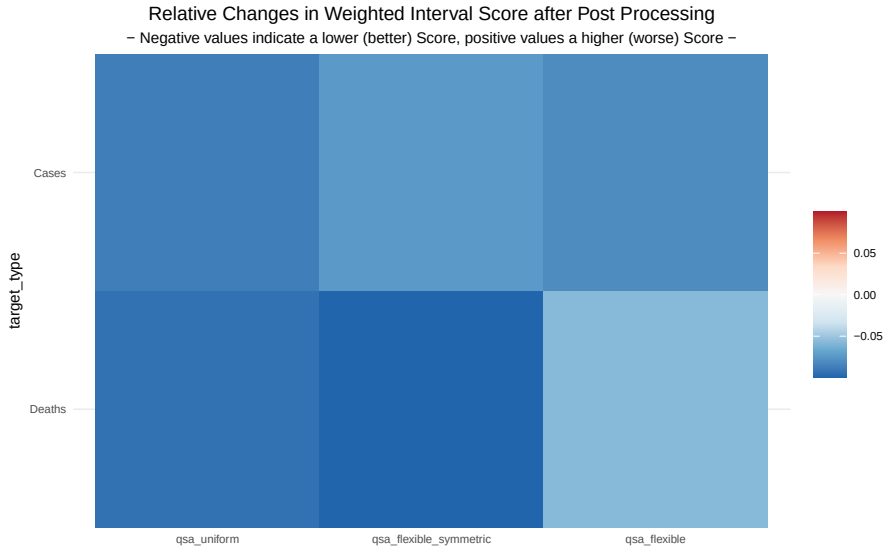

Figure 2: Across both Target Types all QSA mrthods improve the WIS.

If we split models into human and model forecasts the results change as is depicted in Figure 3. 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 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 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.

0.1.4 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 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.

Again, a split of the post processed models into human and model forecasts reveals differences as depicted in **??**. 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.

Additionally breaking down the results by target types as well, reveals the patterns in **??** 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.

0.1.5 Quantiles

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

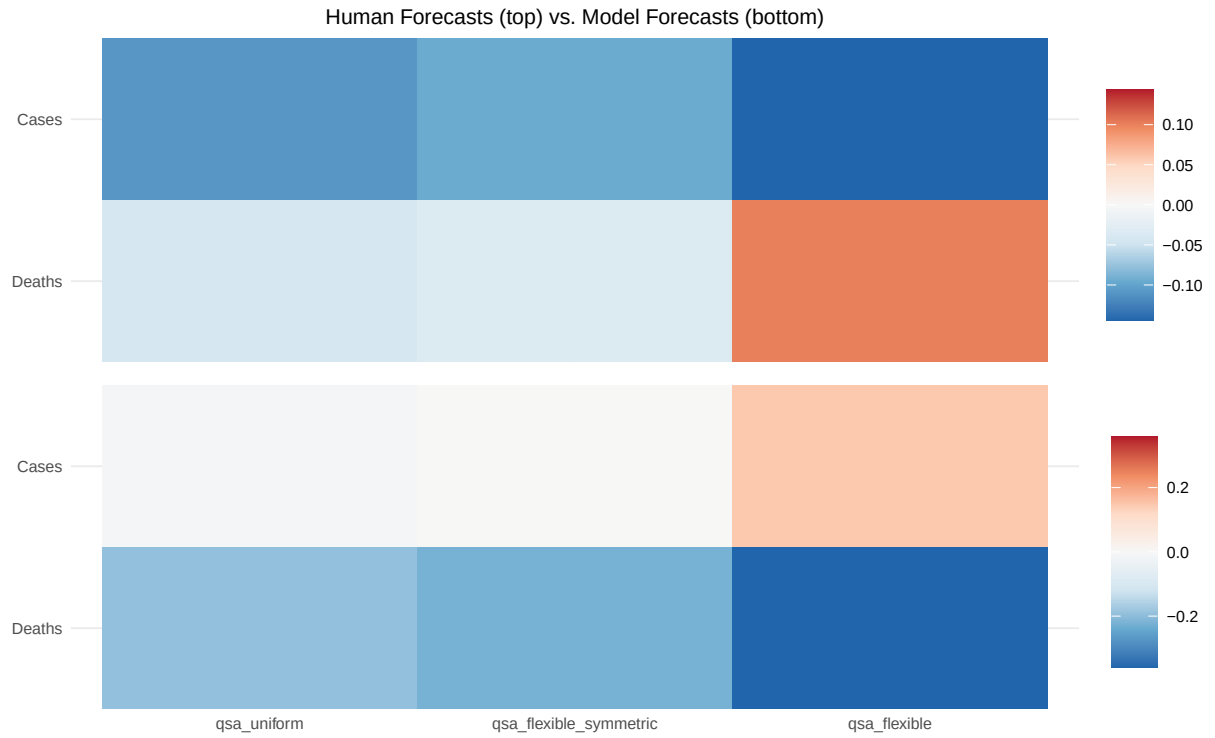

Figure 3: Forecasting improvements differ across target types for different model groups. Human forecasts are primarily improved for cases, while model forecast improvments are only found for deaths.

prediction intervals lying within the 0.25 and 0.75 quantiles near to no improvements can be observed for either flavor. This 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 preform better, particularly for intervals `qsa_uniform` could not improve due to its restrictive nature. Apparently the intervals with coverages equal or smaller than 50 % 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.

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 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 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.

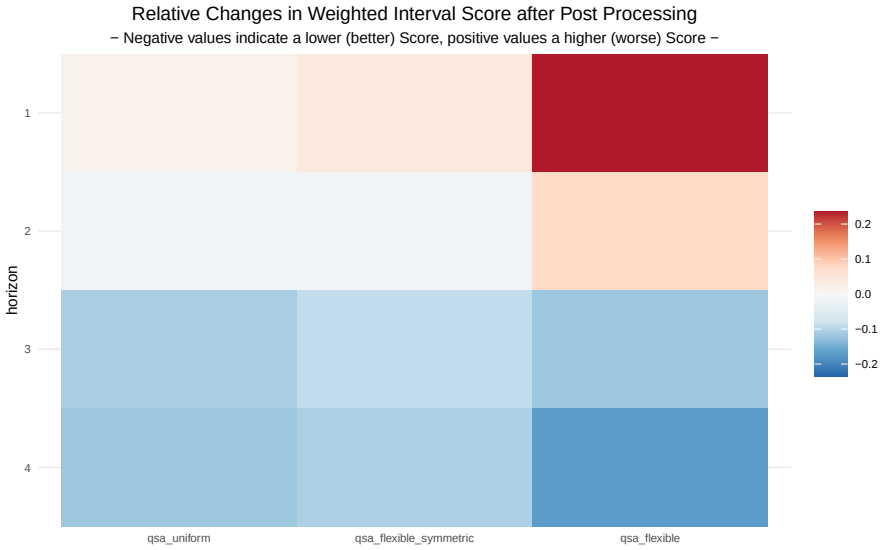

Figure 4: QSA methods improve forecasts more for larger horizons. For smaller horizons they tend to overfit, this is especially the case with QSA Flexible.

0.1.6 Conclusion

Overall `qsa_uniform` is the QSA flavor that preformed best among the UK dataset time series. It produced notable improvements to the WIS in the validation set, without overfitting the training data. The additional flexibility among interval adjustments that `qsa_flexible_symmetric` provided did not bring any notable gains. Most surprisingly it could not improve the WIS for smaller prediction intervals. It rather had the tendency to slightly overfit the data. The additional flexibility of non-symmetric interval adjustments `qsa_flexible` offered, had less clear effects. While overall, as `qsa_flexible` and `qsa_flexible_symmetric` could not lower the WIS more than `qsa_uniform`, `qsa_flexible` did outpreform the others methods in the szenarios where post processing was most useful. `qsa_flexible` did however also substantially overfit 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, however `qsa_flexible` can be a better fit for data requiring large and and varying adjustments across quantiles. In regards for when to use QSA, it preformed best when forecasts underestimated uncertainty which was the case for larger horizons as well as more extreme quantiles. Furthermore, the method performance also depended on the types of models as well as the target type. Both taken separately, QSA worked best for human forecast models and cases. Observed together we did find that QSA preformed best for the human forecasts of cases and model forecasts of deaths. The former findings, are in line the hypothesizes that humans can not grasp uncertainty as well as models and that forecasting uncertainty is higher for cases than for deaths, because deaths are strongly linked to past cases.

Finally, these results leave much room for further investigating and improving the QSA method. For one, it would be interesting to see whether the described results hold for the EU Hub dataset. 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 also be to run `qsa_flexible_symmetric` and `qsa_flexible` with penalization. This would however require keeping some observations as test set as we would fit the penalization strength to the validation set and thus a longer time series as in the EU Hub dataset.

In regards to additional methods, a asymmetric version of `qsa_uniform` whith 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 soly to the asymmetry or are also partially due 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

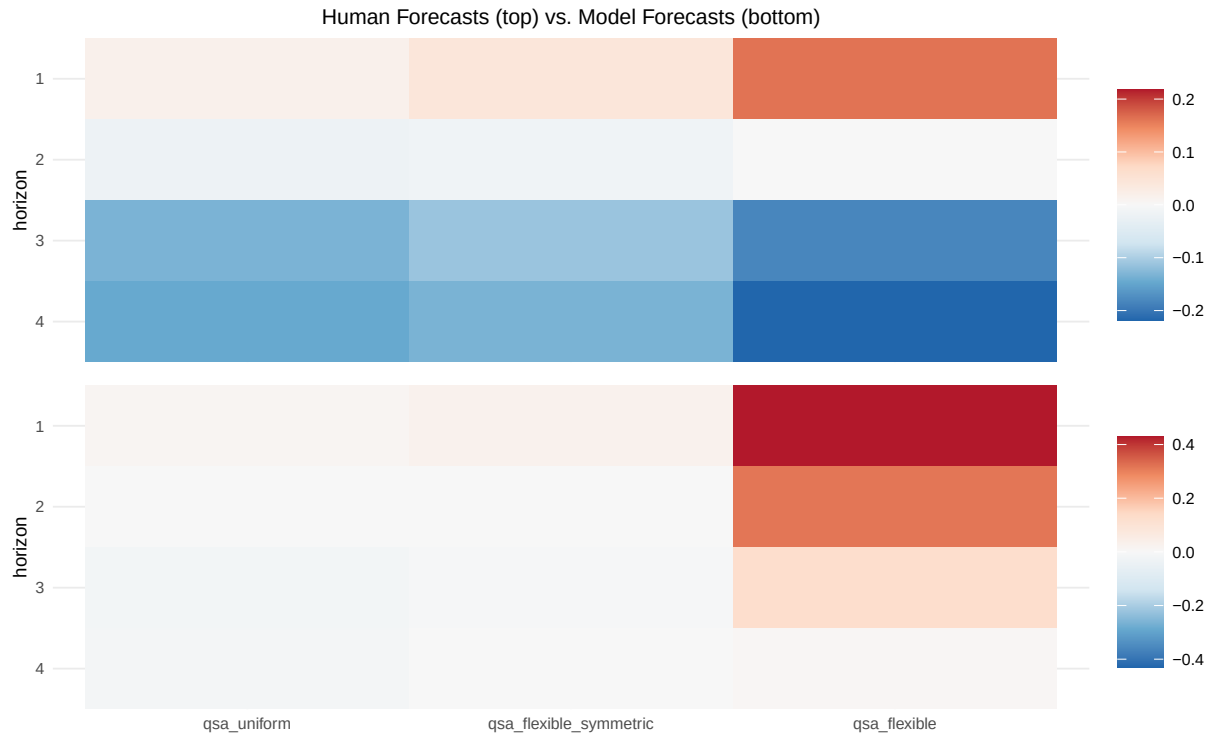

Figure 5: Forecasting improvements differ across horizons for different model groups. Human forecasts are primarily improved for horizons larger than 2, while model forecast are not improved at all and are overfitted with QSA Flexible.

form that penalizes towards no adjustments hence towards all QSA factors being equal to one. A final improvement we would also further investigate for larger datasets is to weight the importance of observations in the optimization by there time point. We suggest an exponential smoothing approach that could weight more recent observations stronger than further past ones to allow the optimization of QSA to adapt faster to changes in the data. This would introduce an additional smoothing hyperparameter that would also require fitting on the validation set as for the penalizations.

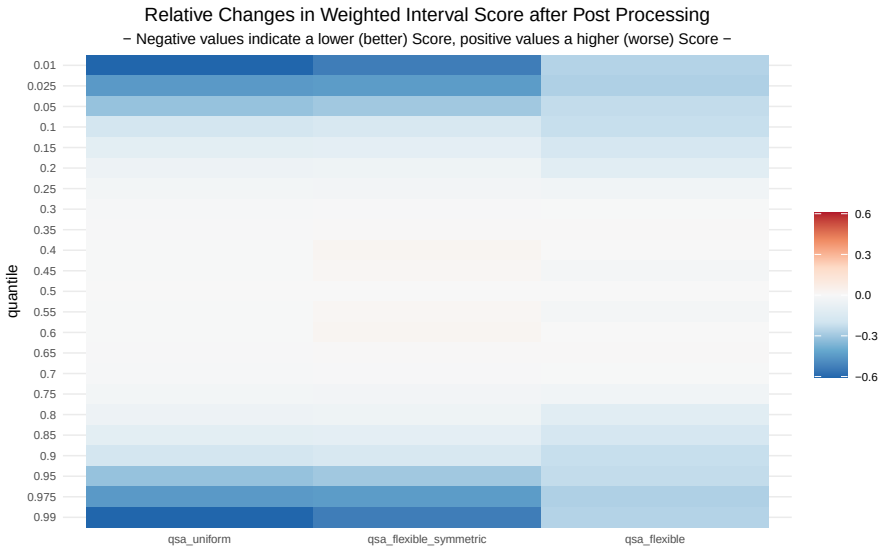

Figure 6: QSA improves forecasters larger for more extreme qunatiles and thus larger intervals.

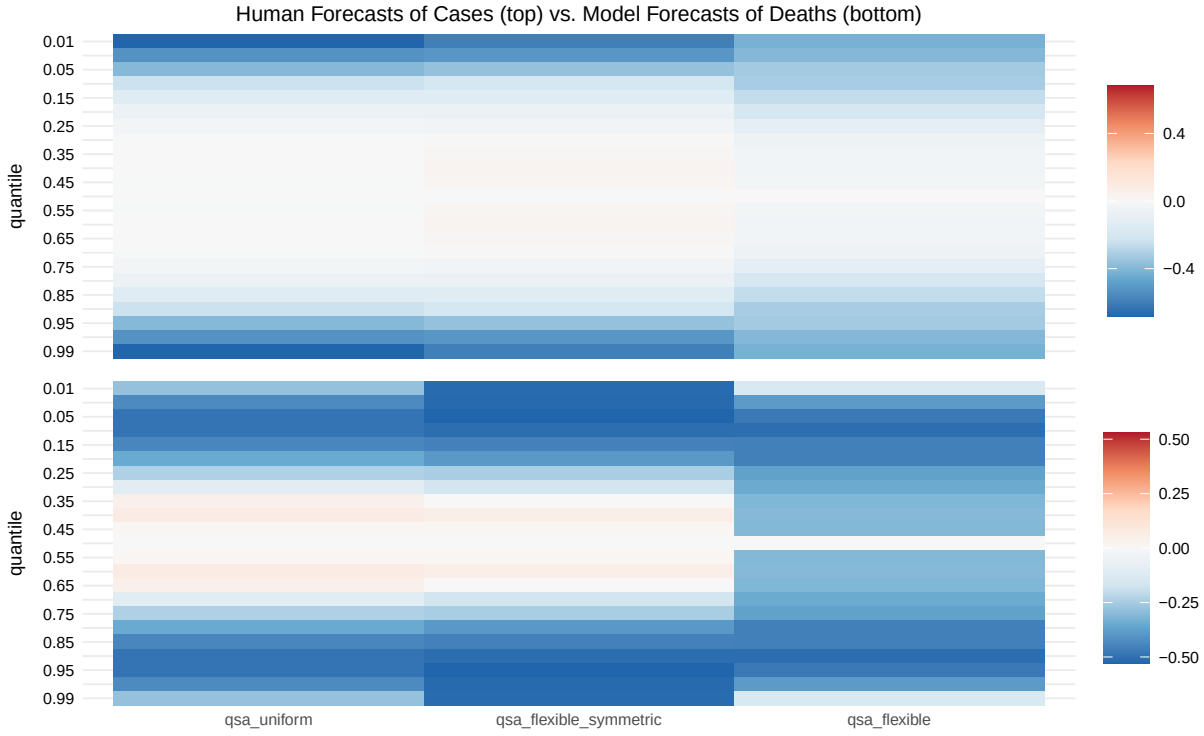

Figure 7: QSA Flexible overfits as it intervals are to low in the training and to high in the validation set.

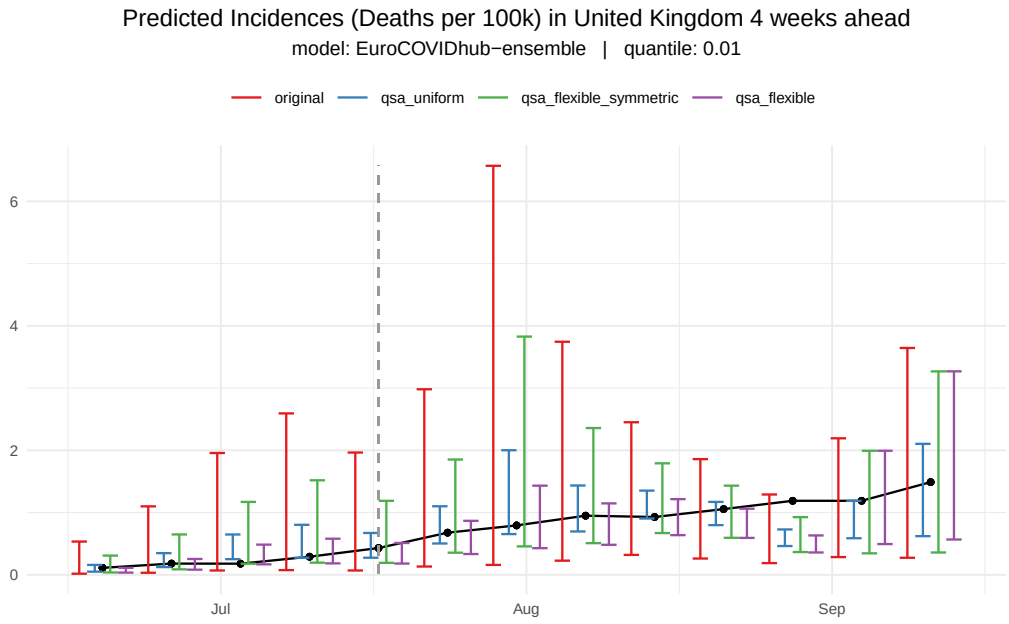

Figure 8: QSA can underestimate uncertainty for extreme quantiles and few data points to learn from.