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

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.

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 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 percent, while the overprediction drops by -18.1834865 which in relativ terms are percent. The main driver behind the increasing in the intervals, is that they do not reach high enough. Thus by increasing the intervals and achieving better coverage of larger 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 especially 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.

\begin{table}

\caption{QSA Flexible_Symmetric improves WIS by increasing interval widths.}

method	interval_score	dispersion	underprediction	overprediction
original	65.74	12.00	5.83	47.91
qsa_flexible_symmetric	60.92	33.22	2.49	25.21

\end{table}

Table Section 0.1 presents the aggregated performance of `qsa_flexible_symmetric` on the validation set.

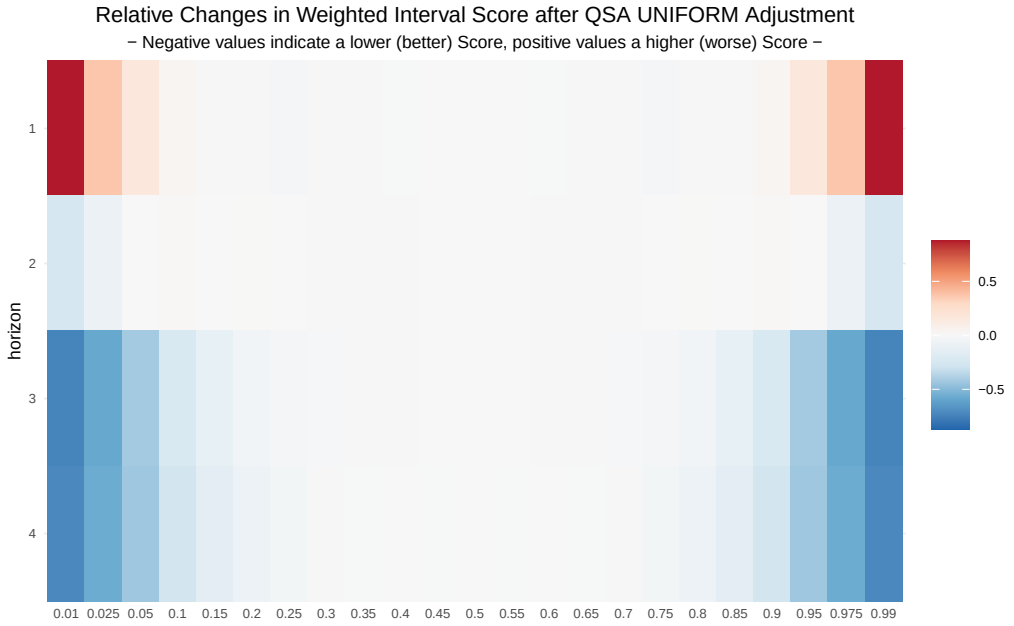

Figure 1: QSA Uniform beneficial for extreme Quantiles at large Horizons.

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 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`.

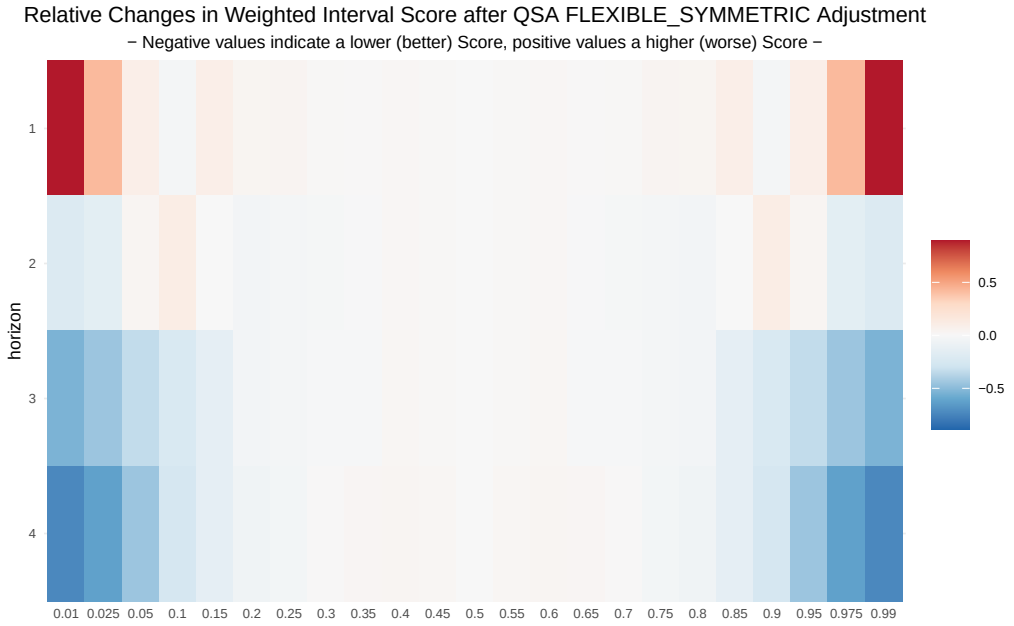

Figure 2: QSA Flexible Symmetric overfits Quantiles at short Horizons.

Table 2: QSA Flexible improves WIS by moving intervals upward.

method	interval_score	dispersion	underprediction	overprediction
original	65.74	12.00	5.83	47.91
qsa_flexible	60.47	25.31	9.31	25.84

Table 2 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 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.