Ensemble Comparison Project

Li Shandross

2023-09-14

Introduction/Background

Overview:

Goal: given quantile forecasts as an input, compare: - linear pools - direct computation of quantile average or median

Related papers:

- By Emily: https://royalsocietypublishing.org/doi/10.1098/rsif.2022.0659
- Busetti: https://onlinelibrary.wiley.com/doi/full/10.1111/obes.12163?casa_token=J9KIeOZWEUAAAAAA%3A8Xm3PyoD9TqAczQf1P6C9QOEfHA-KePb06cLGtlEdM-MhFpbrXIHmD1bdP9tO0gcGiwh4mjPlcXtNVA

Our setting:

- Primary interest is flu hospital admissions: will FluSight use the linear pool method this fall?
- But we only have 2 seasons of quantile forecasts for flu, and in the first season the data was iffy (full reporting only started halfway through the season). Therefore, if possible let's also include COVID hospital admissions

Methods

Start with flu:

- Both seasons
- All locations forecasted in the flusight exercise. Anticipating that there might be a difference, let's formally collect somewhere the locations used in flu so that we can easily use the same ones for covid if/when we add it in later.
- Include all forecast horizons, we can stratify results by horizon afterward
- All component models, no checks like whether the model provides all locations or all horizons. But the model should provide all quantile levels for any location/horizons where it provides any forecasts.

Ensemble methods, all equally weighted:

- Linear pool, using `hubEnsembles::linear_pool`
  - *Note:* we're calling distfromq to get from quantiles to full distributions. It makes some tail assumptions. The default is normal distributions for the tails. This is awkward because hospitalizations have to be non-negative, but a normal lower tail could extend into negative values. To address this, with this method let's truncate any negative quantile estimates to 0.
  - ELR is working on adding in a lognormal tail assumption to distfromq. When it is ready, we can add it into the comparison (so there would be 2 LP methods, LP-normal and LP-lognormal)
- Mean of quantiles (Vincent-mean), using `hubEnsembles::simple_ensemble`
- Median of quantiles (Vincent-median), using `hubEnsembles::simple_ensemble`

Data

Flu incident hospitalization truth and forecasts are queried from zoltar for the 2021-2022 and 2022-23 seasons. This is the only target/time period evaluated currently.

Forecast locations

Forecasts are made for all 50 states, Washington DC, Puerto Rico, the Virgin Islands come and the US as a whole. 1 - 4 week ahead incident flu hospitalizations Only validation phase, no testing phase (assume that will be the upcoming season).

Model specifications

Metrics and evaluation

- different ways to stratify evaluations: overall, by season, by horizon, by location, by (forecast) date
- separate by geographic scale (averaged states/territories vs US national) to avoid the US national results obscuring that of the state/territories
- Incident flu hospitalizations evaluated by usual horizon, incident covid hospitalizations will also be evaluated on a weekly basis
- metrics: average wis, average mae, average 50% pi coverage, average 95% pi coverage, average rwis, average rmae

Results

Plot Forecasts

Will plot forecasts for all models, select locations (including US, highest count state, lowest count state/territory)

Overall model performance

Table 1: Summary of overall model performance across both seasons, averaged over the US national geographic scale.

model	wis	mae	cov50	cov95	rwis	rmae
median-ensemble	620.861	927.605	0.726	0.967	0.682	0.832
lp-normal	740.555	965.316	0.797	0.991	0.813	0.866
lp-lognormal	740.605	965.346	0.797	0.991	0.813	0.866
Flusight-baseline	910.836	1114.453	0.580	0.849	1.000	1.000
mean-ensemble	981.881	1064.878	0.693	0.962	1.078	0.956

Table 2: Summary of overall model performance across both seasons, averaged over the states geographic scale.

model	wis	mae	cov50	cov95	rwis	rmae
median-ensemble	18.158	27.360	0.597	0.922	0.794	0.933
lp-normal	19.745	27.932	0.709	0.990	0.863	0.953
lp-lognormal	19.747	27.933	0.708	0.990	0.863	0.953
mean-ensemble	20.180	29.582	0.595	0.889	0.882	1.009
Flusight-baseline	22.876	29.315	0.604	0.881	1.000	1.000

By season model performance

Table 3: Summary of by season model performance, averaged over the US national geographic scale.

model	season	wis	mae	cov50	cov95	rwis	rmae
median-ensemble	2021-2022	180.395	234.472	0.693	0.989	0.843	0.718
Flusight-baseline	2021-2022	213.998	326.591	0.625	1.000	1.000	1.000
mean-ensemble	2021-2022	214.566	292.673	0.636	1.000	1.003	0.896
lp-normal	2021-2022	253.189	267.105	0.773	1.000	1.183	0.818
lp-lognormal	2021-2022	253.193	267.098	0.773	1.000	1.183	0.818
median-ensemble	2022-2023	933.449	1419.505	0.750	0.952	0.664	0.848
lp-normal	2022-2023	1086.427	1460.821	0.815	0.984	0.773	0.873
lp-lognormal	2022-2023	1086.511	1460.878	0.815	0.984	0.773	0.873
Flusight-baseline	2022-2023	1405.366	1673.581	0.548	0.742	1.000	1.000
mean-ensemble	2022-2023	1526.427	1612.895	0.734	0.935	1.086	0.964

Table 4: Summary of by season model performance, averaged over the states geographic scale.

model	season	wis	mae	cov50	cov95	rwis	rmae
median-ensemble	2021-2022	8.038	12.194	0.570	0.926	0.845	0.908
mean-ensemble	2021-2022	8.398	12.540	0.551	0.896	0.882	0.934
lp-normal	2021-2022	9.074	12.466	0.691	0.993	0.953	0.928
lp-lognormal	2021-2022	9.075	12.466	0.690	0.993	0.953	0.928
Flusight-baseline	2021-2022	9.518	13.433	0.650	0.922	1.000	1.000
median-ensemble	2022-2023	25.341	38.124	0.617	0.919	0.783	0.939
lp-normal	2022-2023	27.317	38.909	0.721	0.988	0.844	0.959
lp-lognormal	2022-2023	27.320	38.909	0.721	0.988	0.844	0.959
mean-ensemble	2022-2023	28.541	41.676	0.626	0.885	0.882	1.027
Flusight-baseline	2022-2023	32.356	40.586	0.572	0.851	1.000	1.000

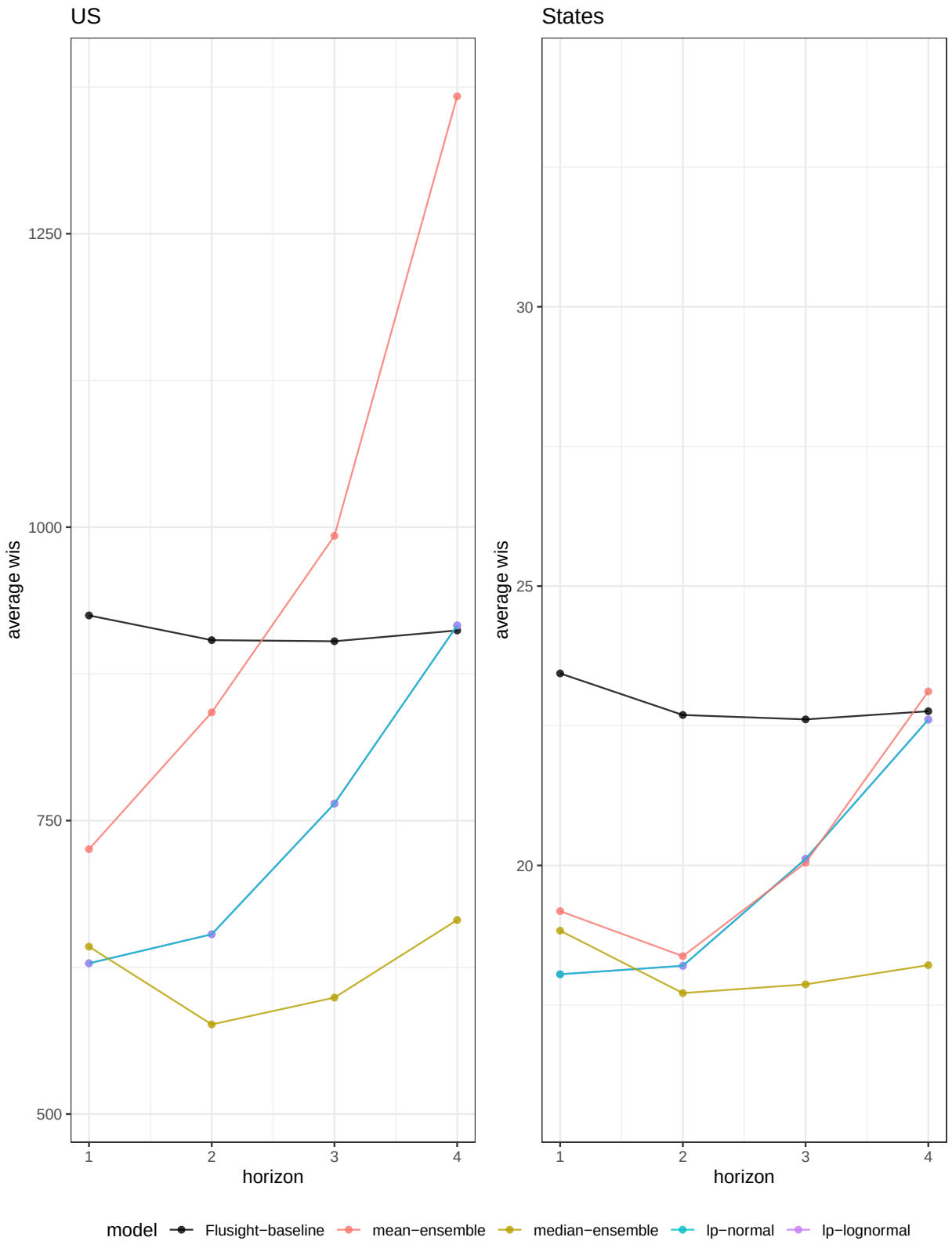
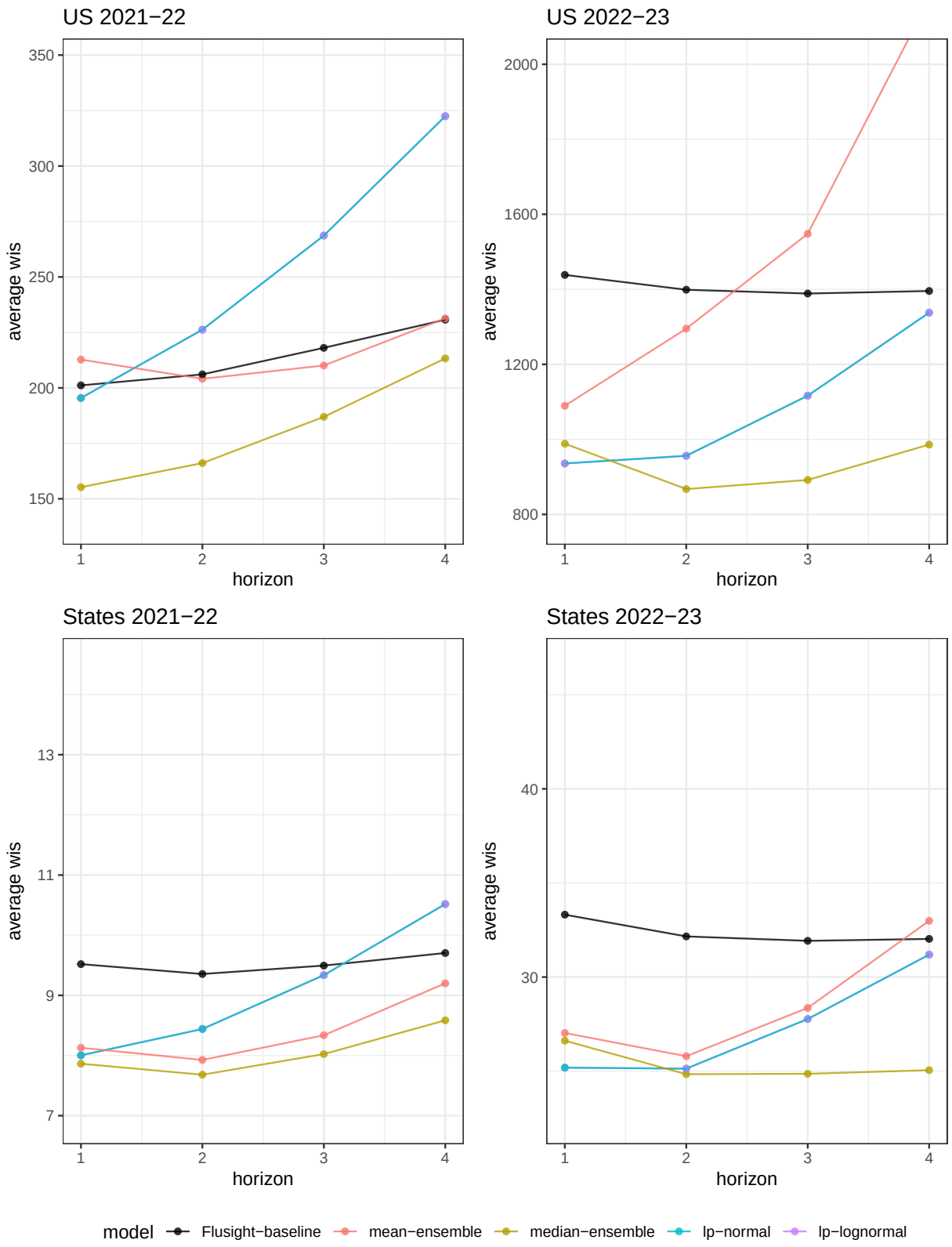


Figure 1: Average WIS by the horizon for each model for the US national level and the states level.

Model performance by horizon

Additional by season split



Model Performance by Week

Model performance by location

Conclusions/Discussion

Overall, the vincent median ensemble is the best of the ensembles. However, there are several instances where the two linear pools (and occasionally the vincent mean ensemble) beat the vincent median; namely, the 1-week ahead horizon for both geographic scales (though it's more pronounced for the averaged states level). A more granular breakdown reveals that this deviation from the overall trend occurs mostly during the 2022-2023 season, specifically during times of high change. These patterns in which the vincent mean ensemble generally performs best, except for the one week ahead horizon with combined seasons or just the 2022-2023 season when the linear pools perform slightly better.

There is only a marginal difference in performance between the two linear pools, even for the week-by-week break down of results. This suggests at least one of three scenarios: 1) the common quantile interpretation from monotonic splines plays a bigger role in defining the shape of the resulting probabilistic distribution the extrapolation for the tails, 2) there is not a meaningful difference between coercing negative values to zero for the linear pool with normal tails and using a linear pool with lognormal tails that are naturally non-negative, or 3) the original forecasts (which are non-negative) ensemble together in such a way that there is negligible difference between the two linear pools, but this may not always be the case for a different set of forecasts.

If the CDC is mainly interested in communicating forecasts with short-term horizons (e.g. 1-2 week ahead), a linear pool ensemble may be worth further investigation, especially during periods of rapid change. Either type of tail distribution would be acceptable, though normal tails with negative values coerced to zero may be preferred simply due to shorter computation time.

References

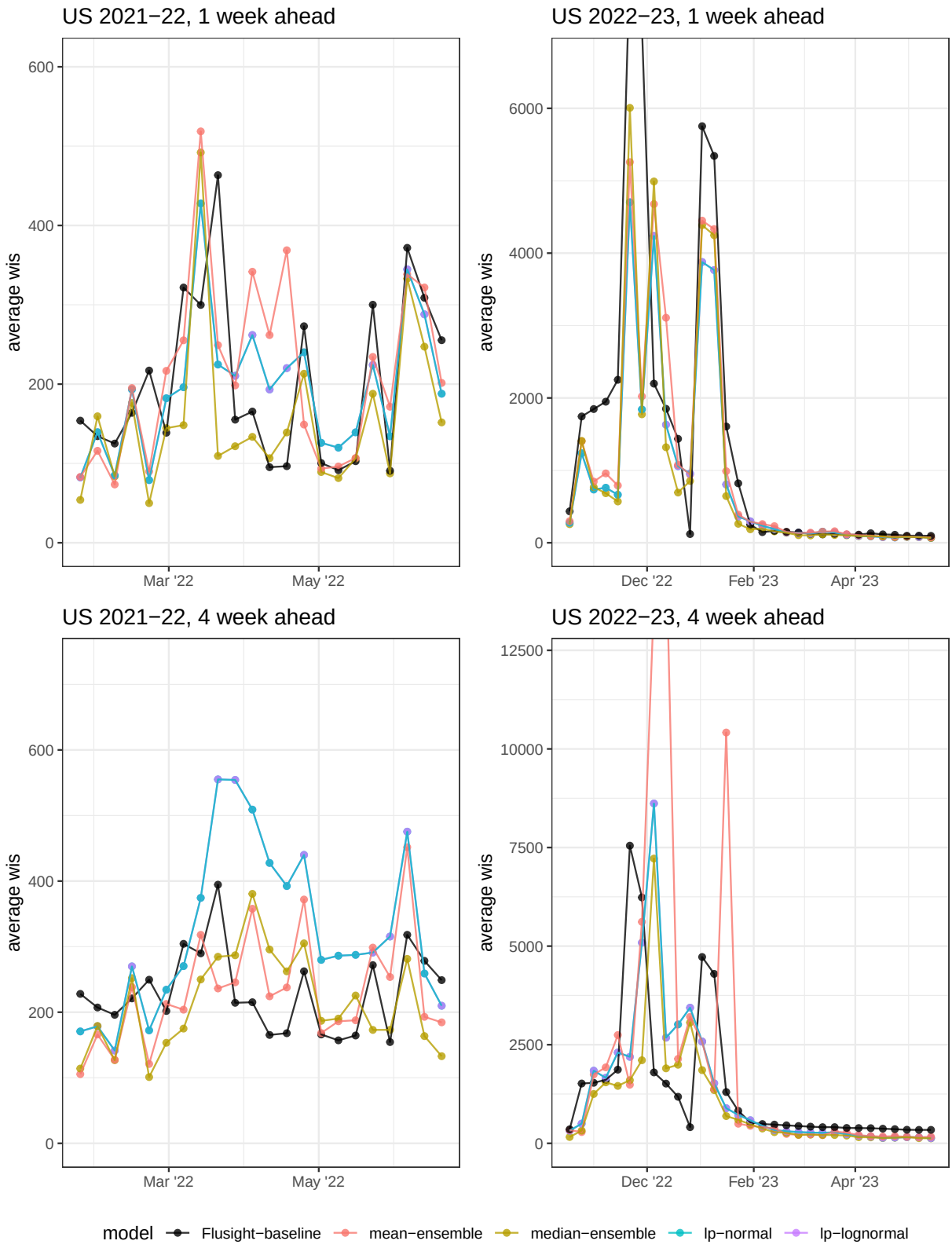


Figure 2: Average h-week ahead WIS and 95% PI coverage for each model for the US national level.

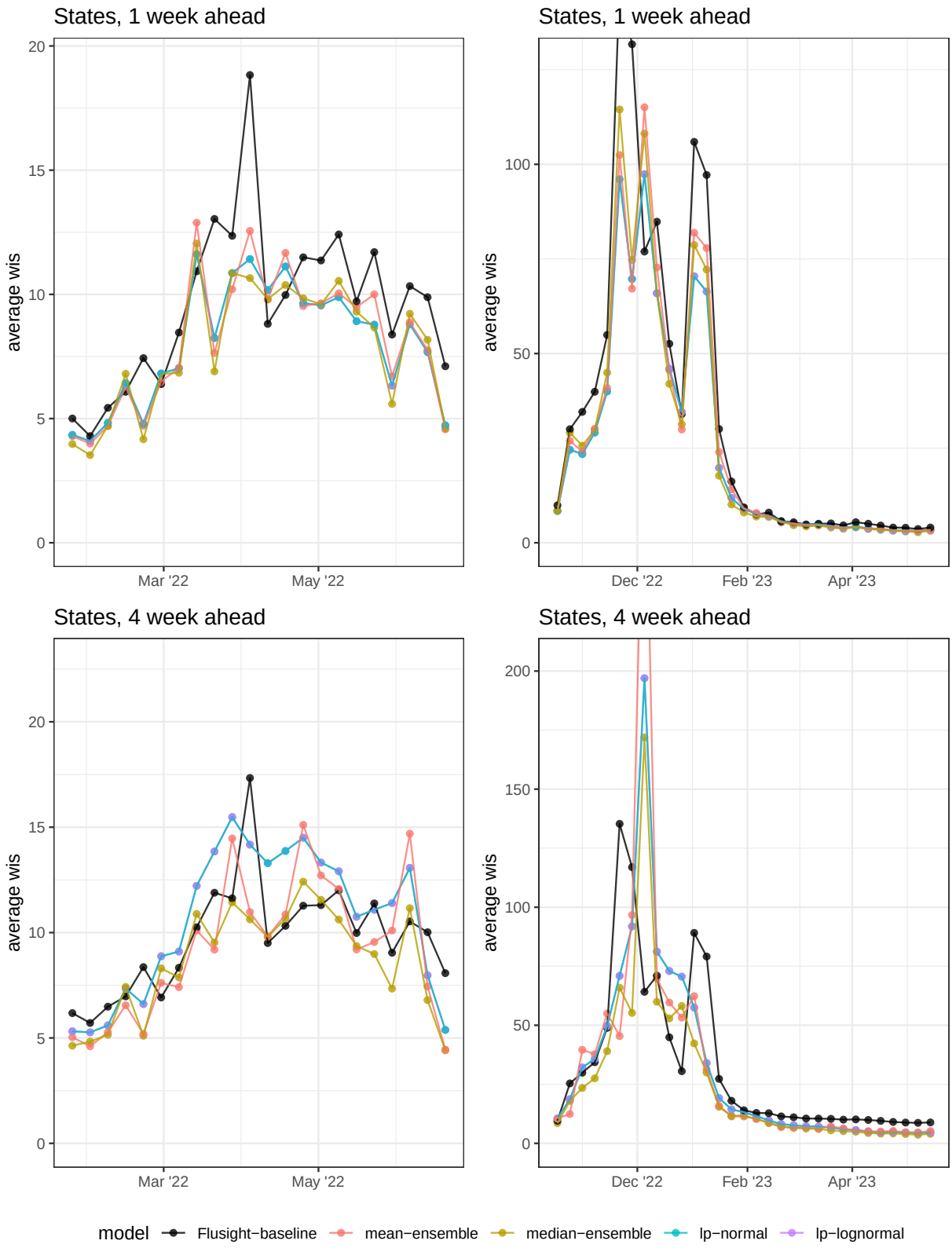


Figure 3: Average h-week ahead WIS and 95% PI coverage for each model for the States national level.

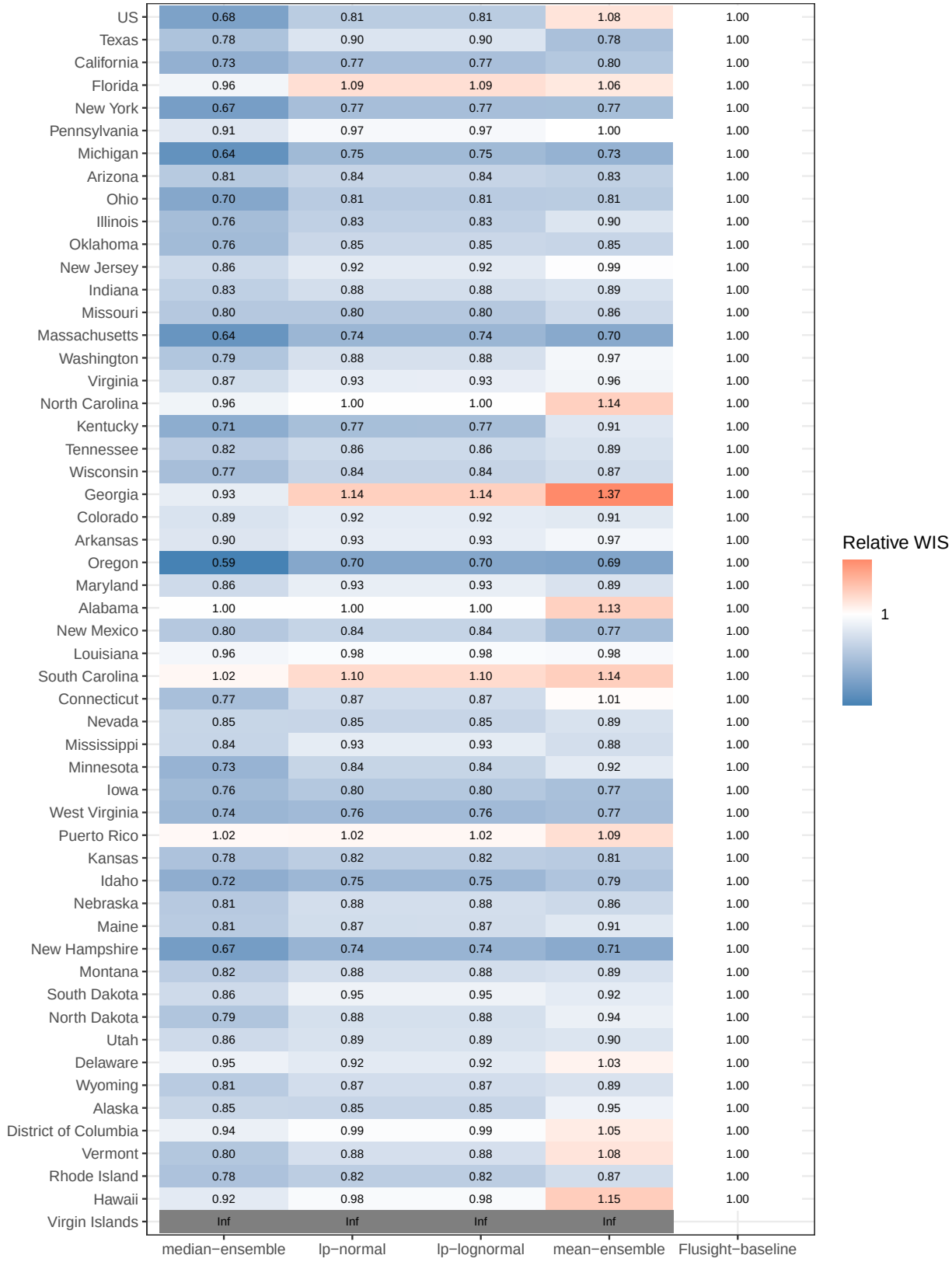


Figure 4: Relative WIS plotted by location for each model across all horizons for both seasons

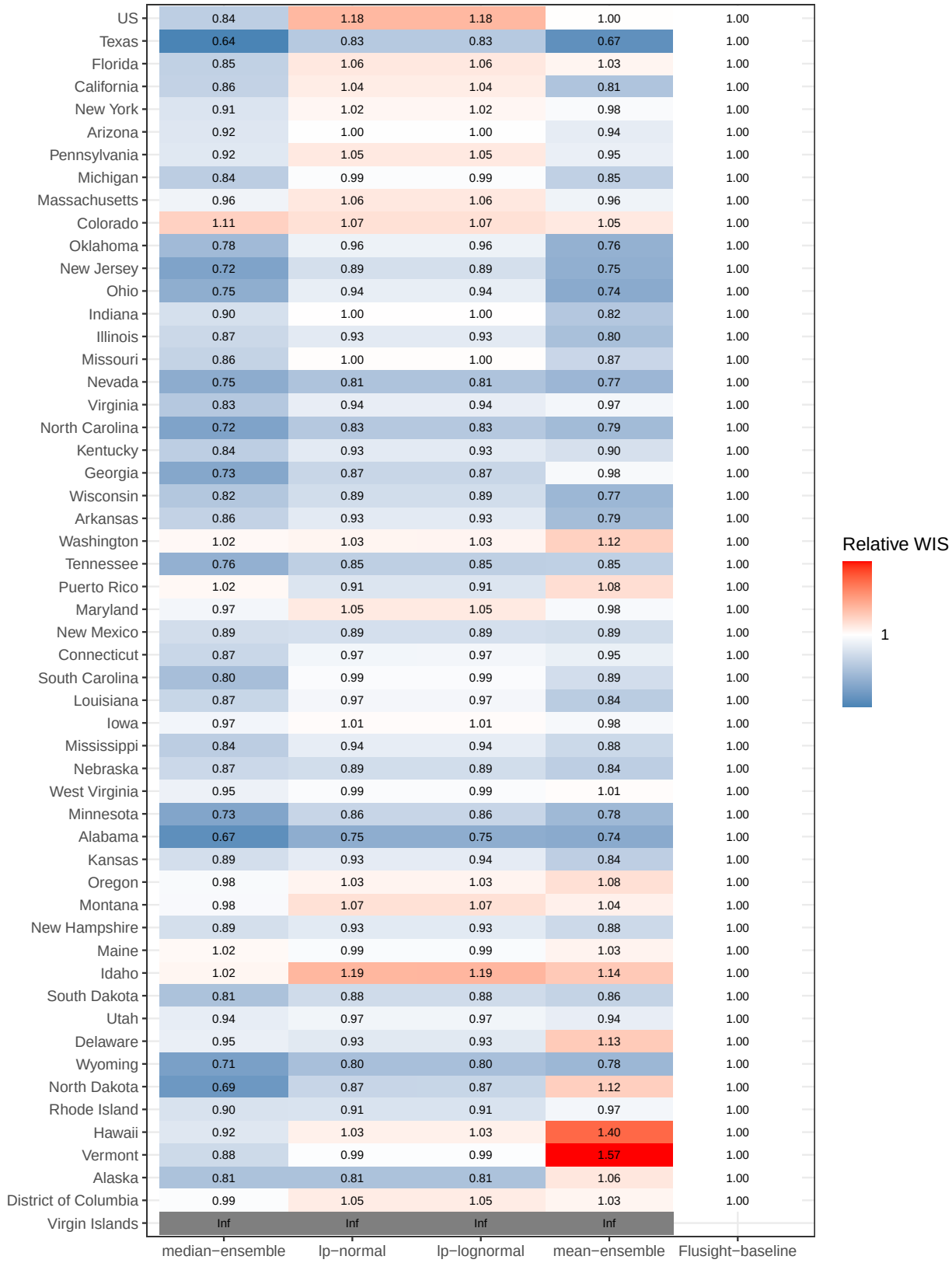


Figure 5: Relative WIS plotted by location for each model across all horizons during the 2021-2022 season

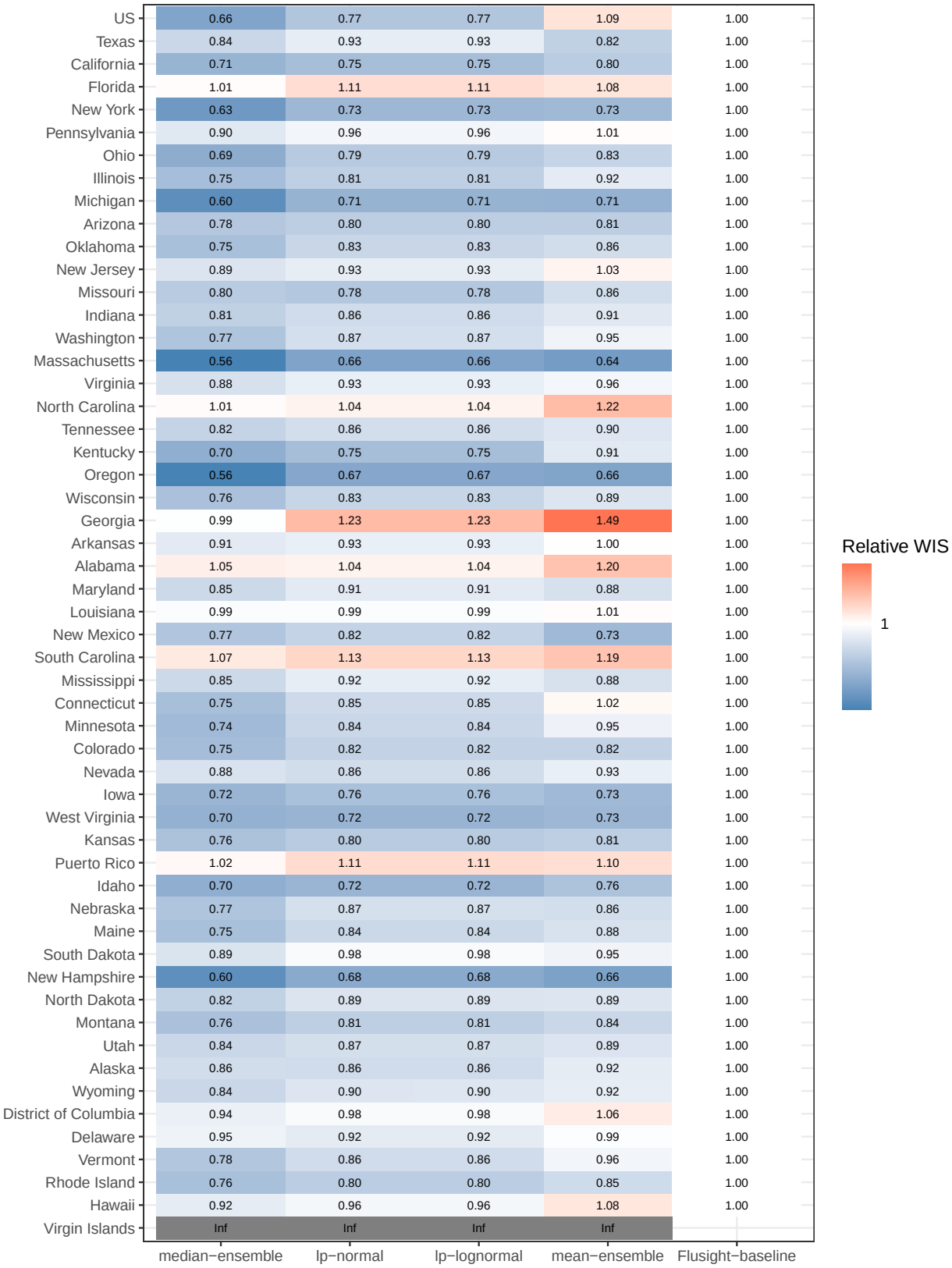


Figure 6: Relative WIS plotted by location for each model across all horizons during the 2022-2023 season

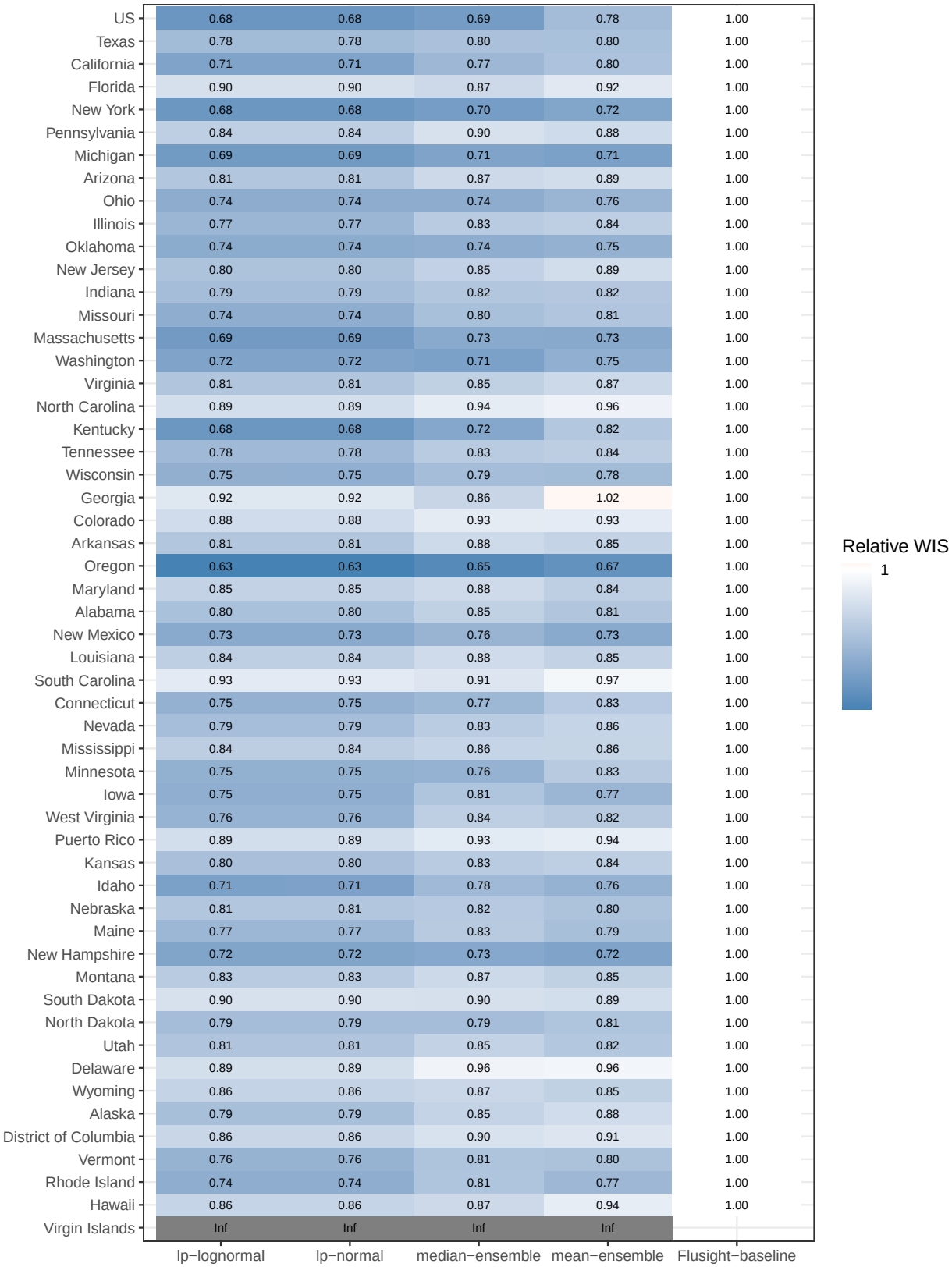


Figure 7: Relative WIS plotted by location for each model across both seasons for a 1 week ahead horizon

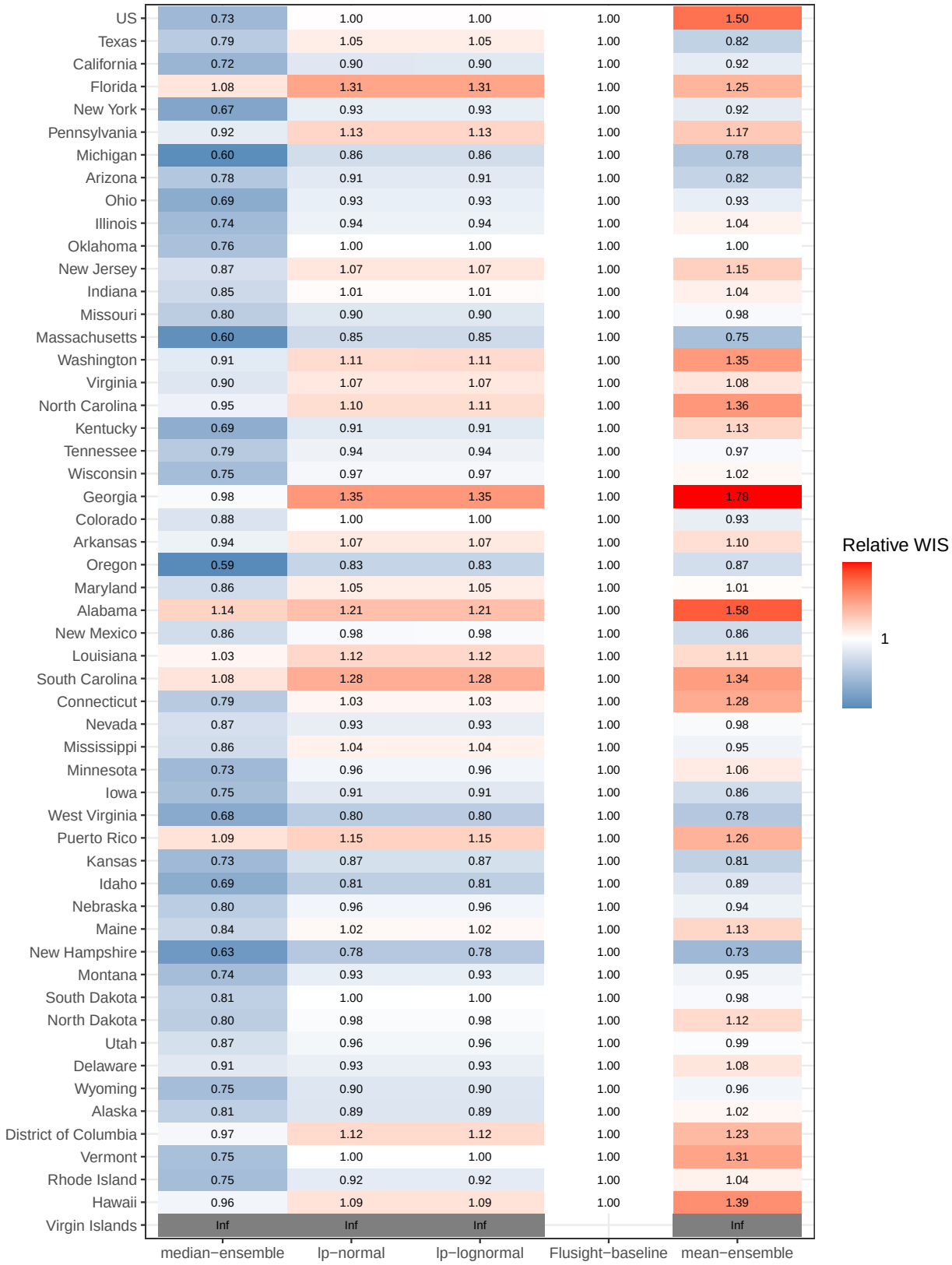


Figure 8: Relative WIS plotted by location for each model across both seasons for a 4 week ahead horizon

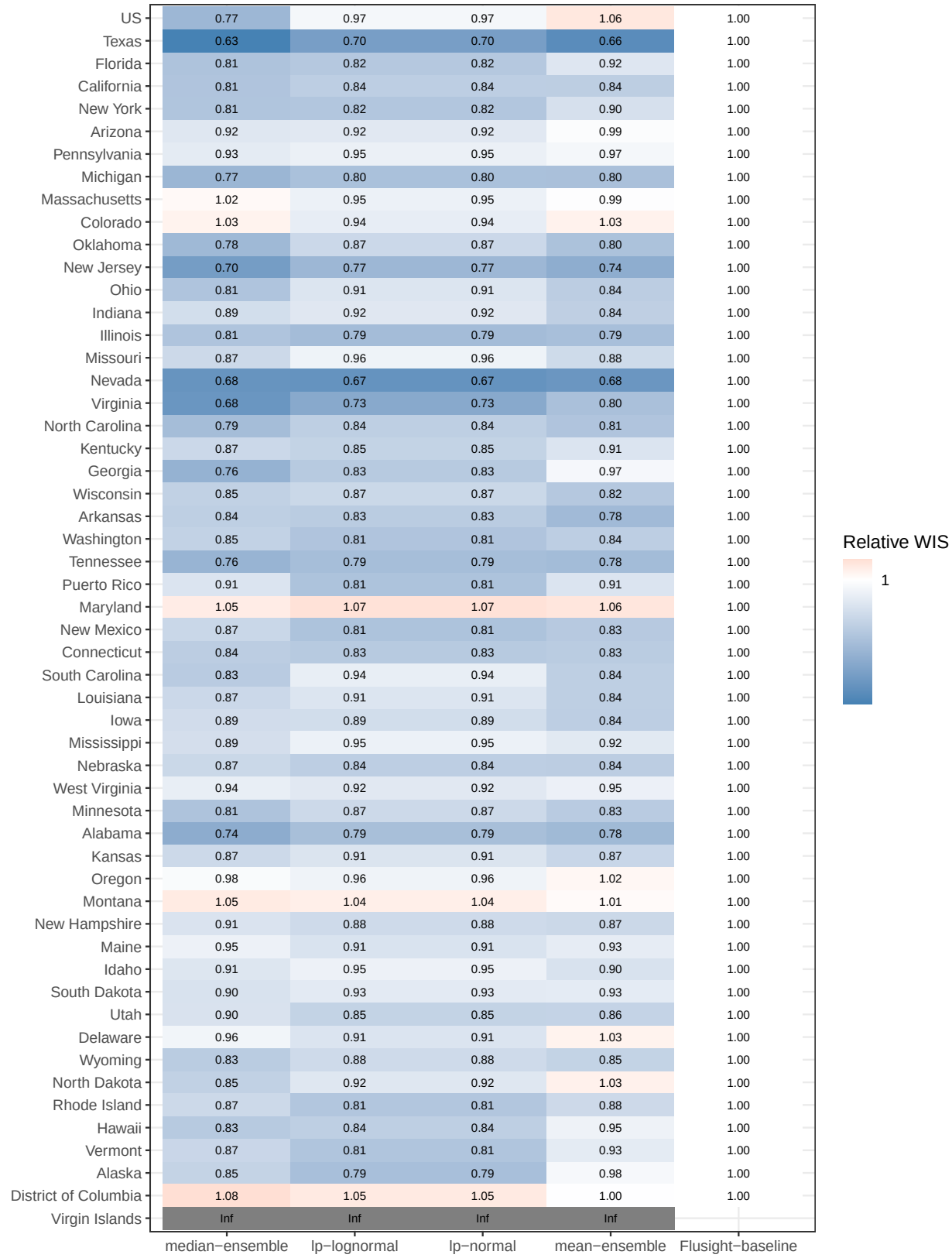


Figure 9: Relative WIS plotted by location for each model for a 1 week ahead horizon during the 2021-2022 season

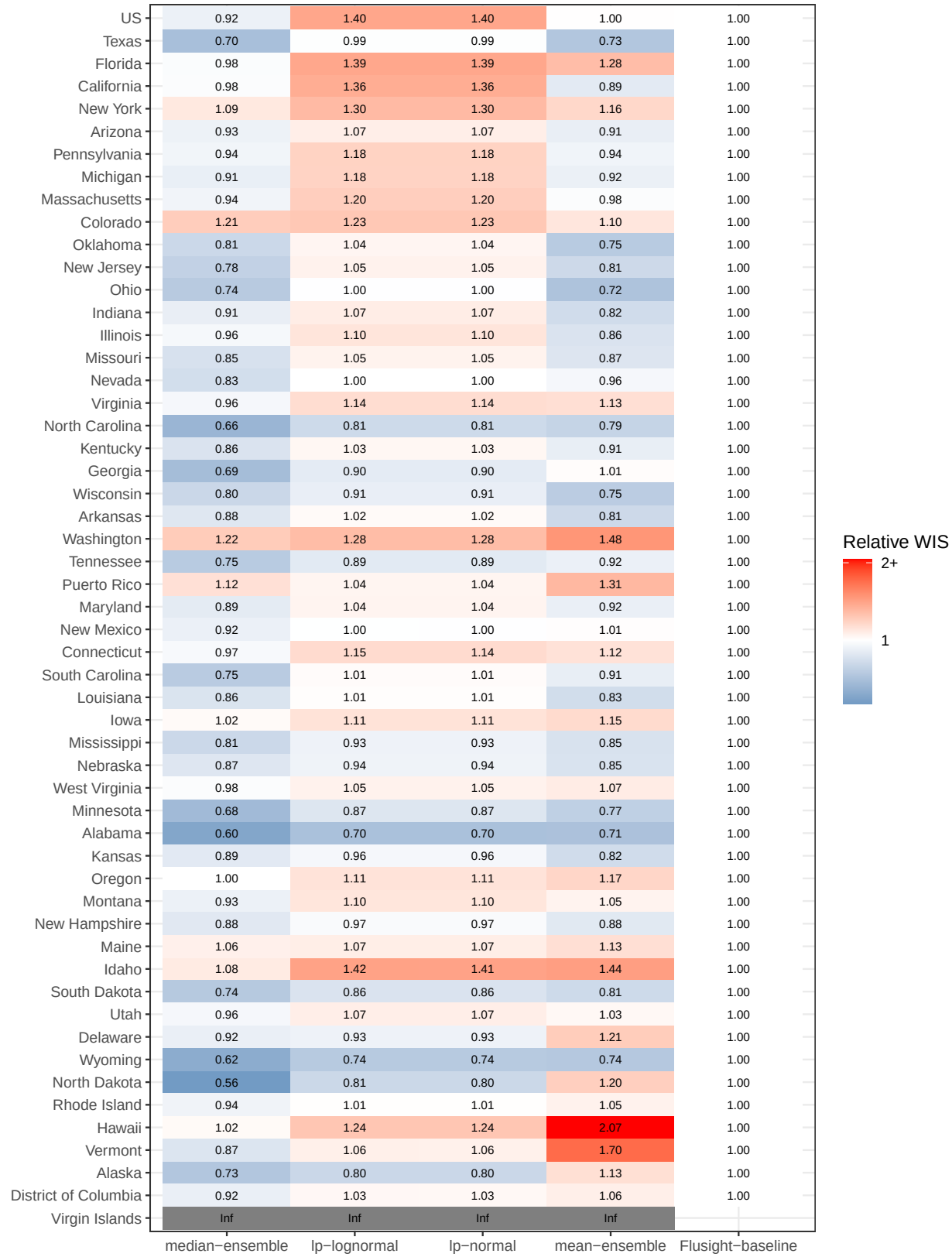


Figure 10: Relative WIS plotted by location for each model for a 4 week ahead horizon during the 2021-2022 season

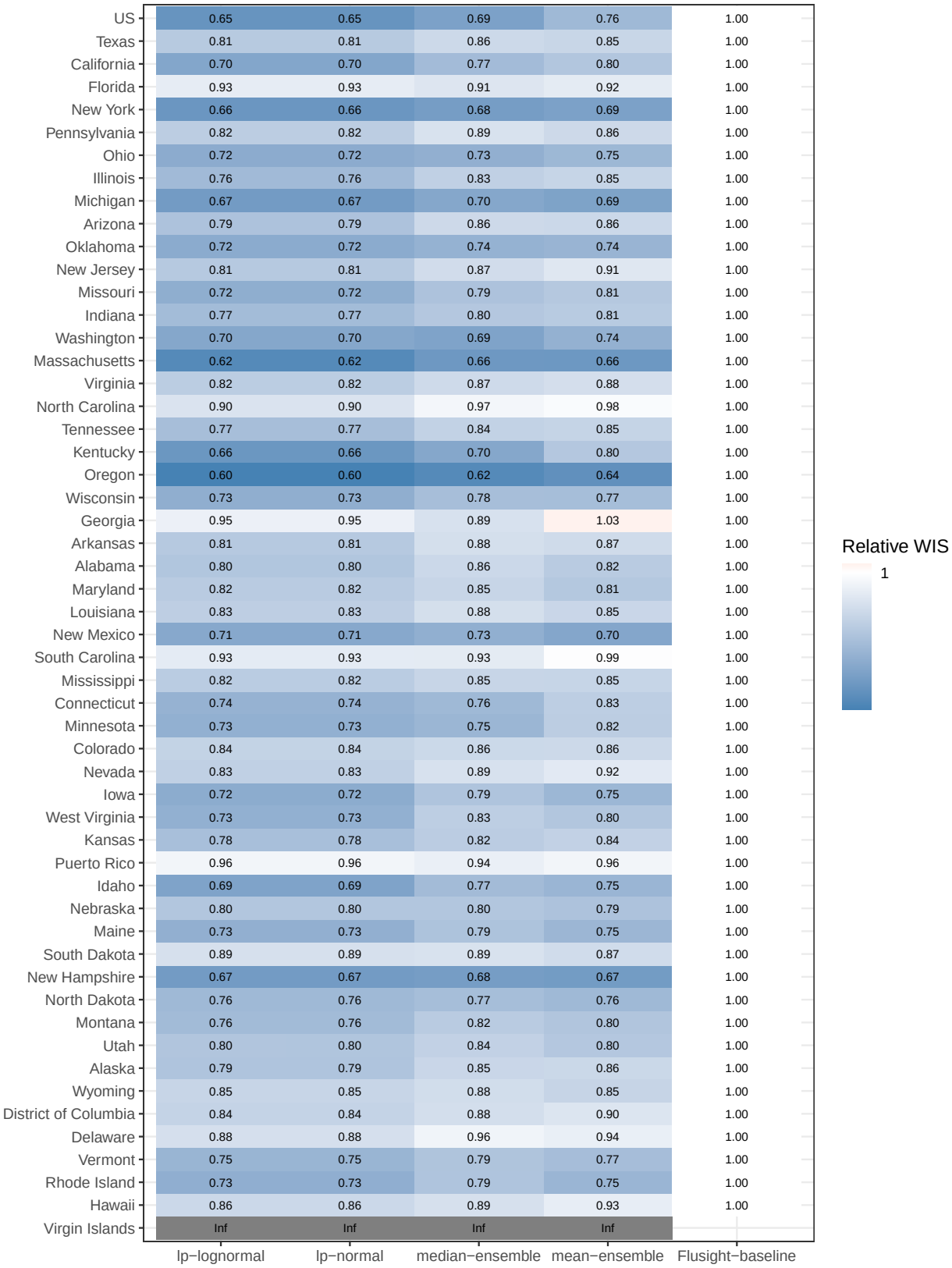


Figure 11: Relative WIS plotted by location for each model for a one week ahead horizon during the 2022-2023 season

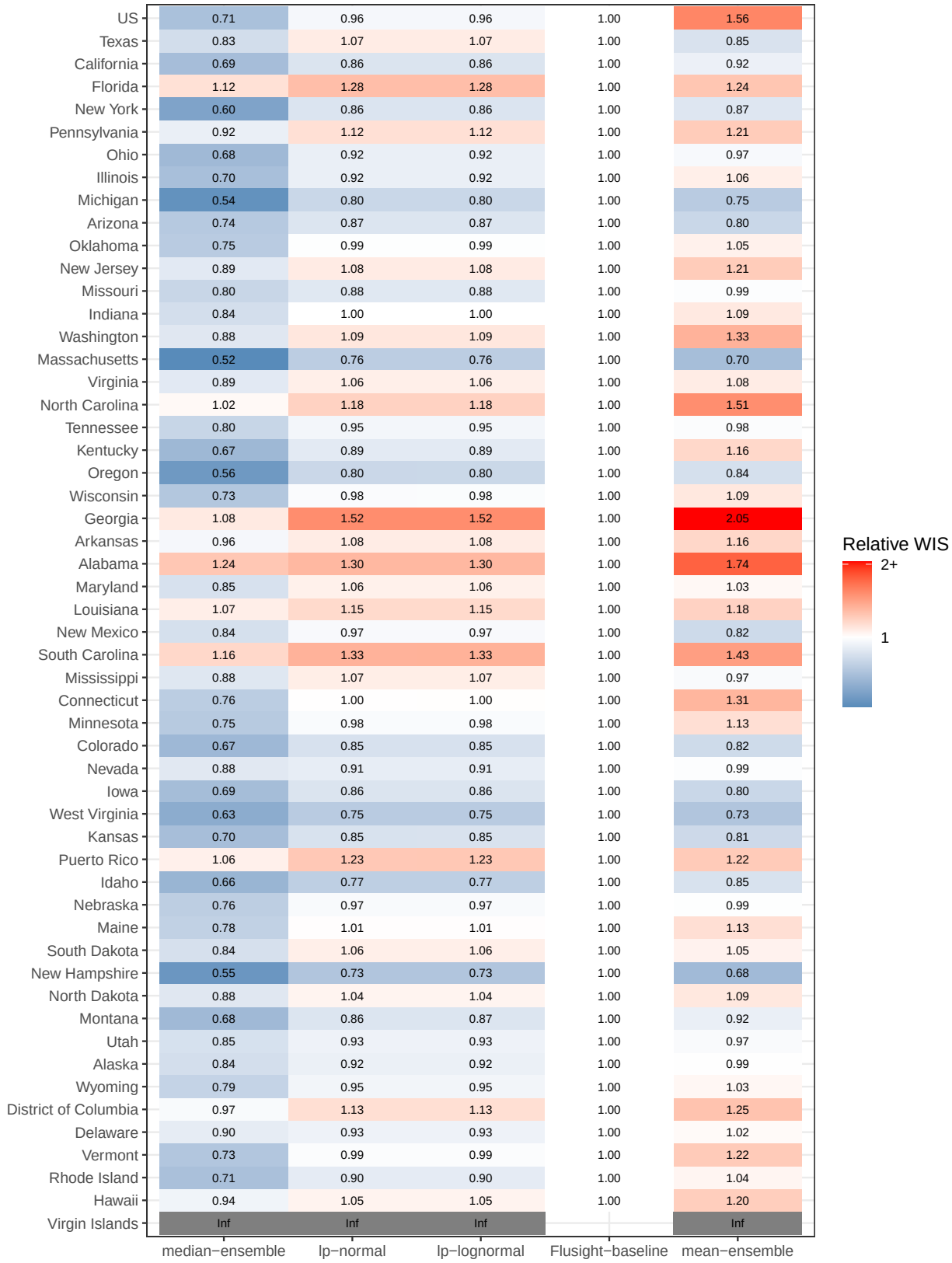


Figure 12: Relative WIS plotted by location for each model for a one week ahead horizon during the 2022-2023 season