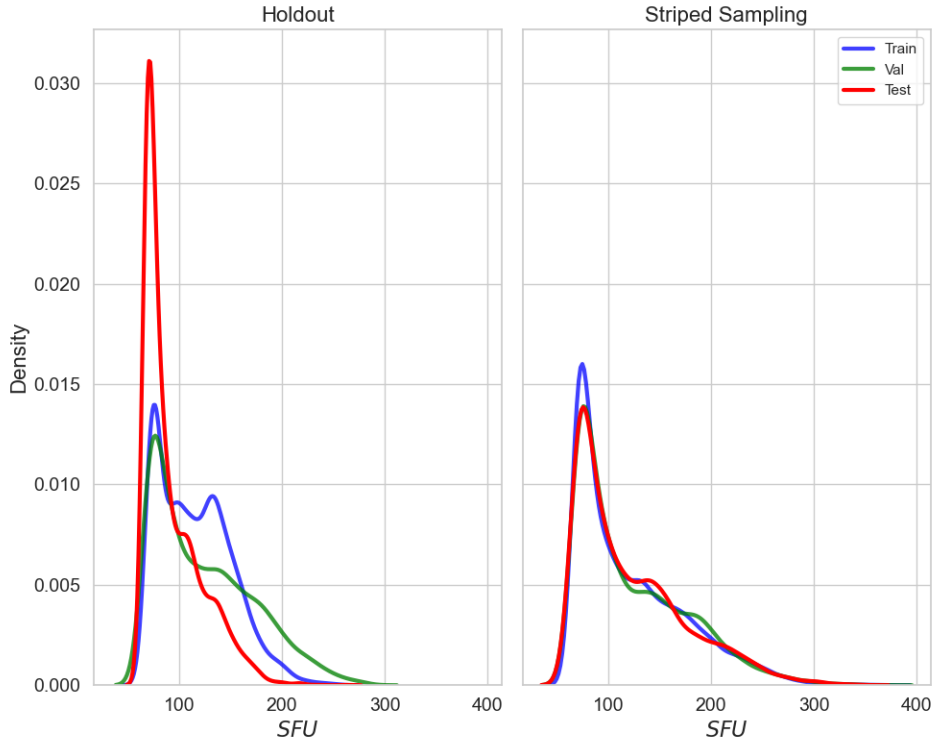
Striped Sampling Compared to Holdout for F10

Joshua Daniell – 12/12/2023

Using the original data from the UV-MLE paper, F10 data from 1947-2021 was considered. Originally, the results showed promise using the standard holdout data sampling method. However, when considered for multivariate, there was a substantial bias, therefore a method known as striped sampling was created.

Striped sampling involves taking the full dataset and splitting it into weeklong segments, these weeklong segments are used sequentially to create "stripes" and can be used to train the model. For the multivariate case, striped sampling showed a substantial improvement and allowed us to apply ML to three newer solar drivers. After these improvements were seen, it was desired to go back and determine if that same striped sampling would improve the UV-MLE F10 work from the original Univariate paper.

Statistical Analysis



In the multivariate paper, we saw that striped sampling created statistically similar datasets. This is still the case when applied to F10, and creates even better sets for ML applications (less biased than holdout).

In the original univariate paper, we used holdout to be consistent with previous work (NBEATS). We saw a difference in performance between the 3 sets, and by analyzing their distributions it is clear that the test set in particular is mostly low solar activity, leading to a bias in predictions. Models were trained, validated, and tested on statistically dissimilar sets, leading to inconsistent performance. Models trained with the striped sampled data *SHOULD* do better in general.

Metrics

Univariate LSTM models with varied lookback were trained and used for prediction, being combined with average of prediction value. The metrics are relative metrics which have been scaled against the persistence model. A value of 1 indicates the same performance as persistence and RMSE and MAPE less than 1 indicate an improvement, a correlation coefficient greater than 1 is also an improvement.

Method	Relative Metric	Train	Val	Test
Holdout*	**RMSE**	0.723	0.815	0.637
	MAPE	0.877	0.822	0.722
	R	1.071	1.024	1.043
Striped Sampling	**RMSE**	0.459	0.463	0.443
	MAPE	0.445	0.452	0.448
	R	1.113	1.121	1.128

Percent Improvement			
Relative Metric	Train	Val	Test
RMSE	37%	43%	30%
MAPE	49%	45%	38%
R	4%	9%	8%

* Uses original Holdout set from UV-MLE Paper

We see pretty dramatic improvements in all cases. It is clear that there is a benefit from using striped sampled data instead of holdout. It should be noted that these datasets are not directly compared, as we cannot enforce the two sets to produce the same training, validation, and test sets. These 3 sets are not the same, but should still provide a general idea of how “good” a model is doing.

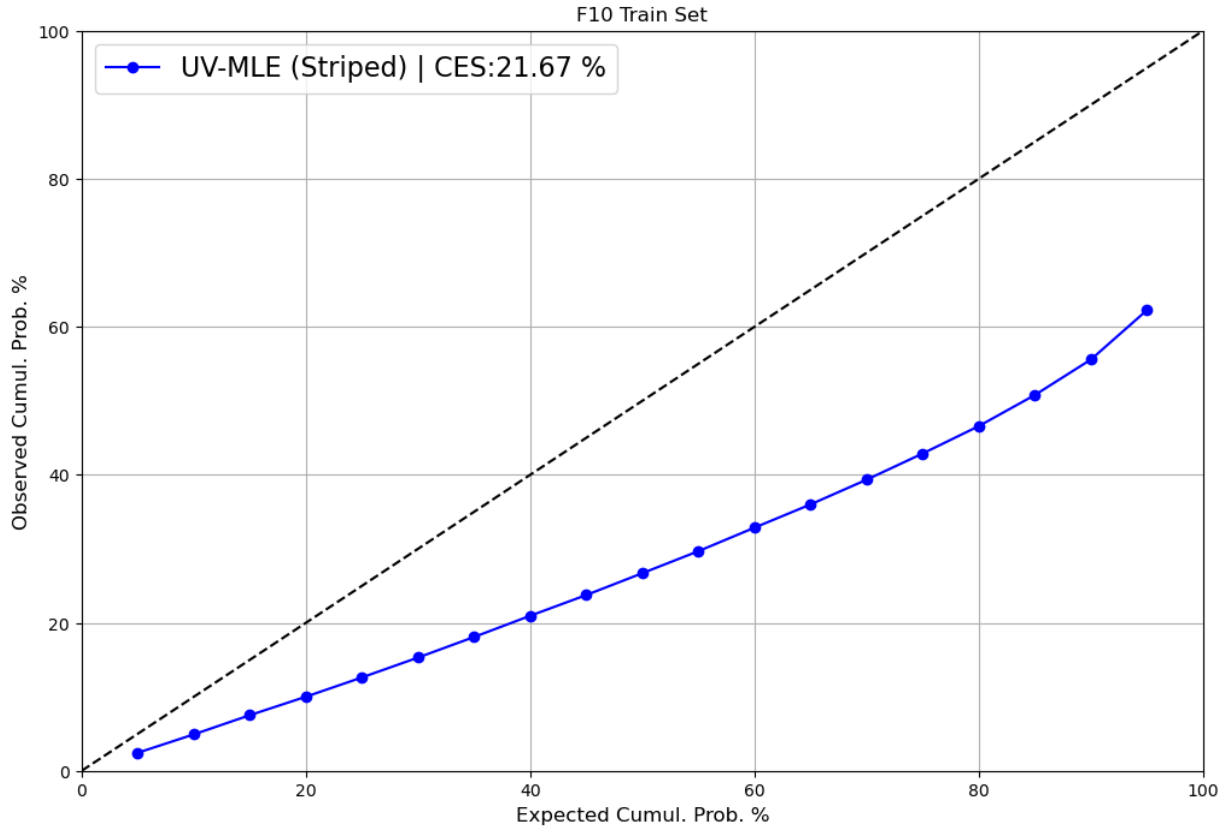
Uncertainty Quantification

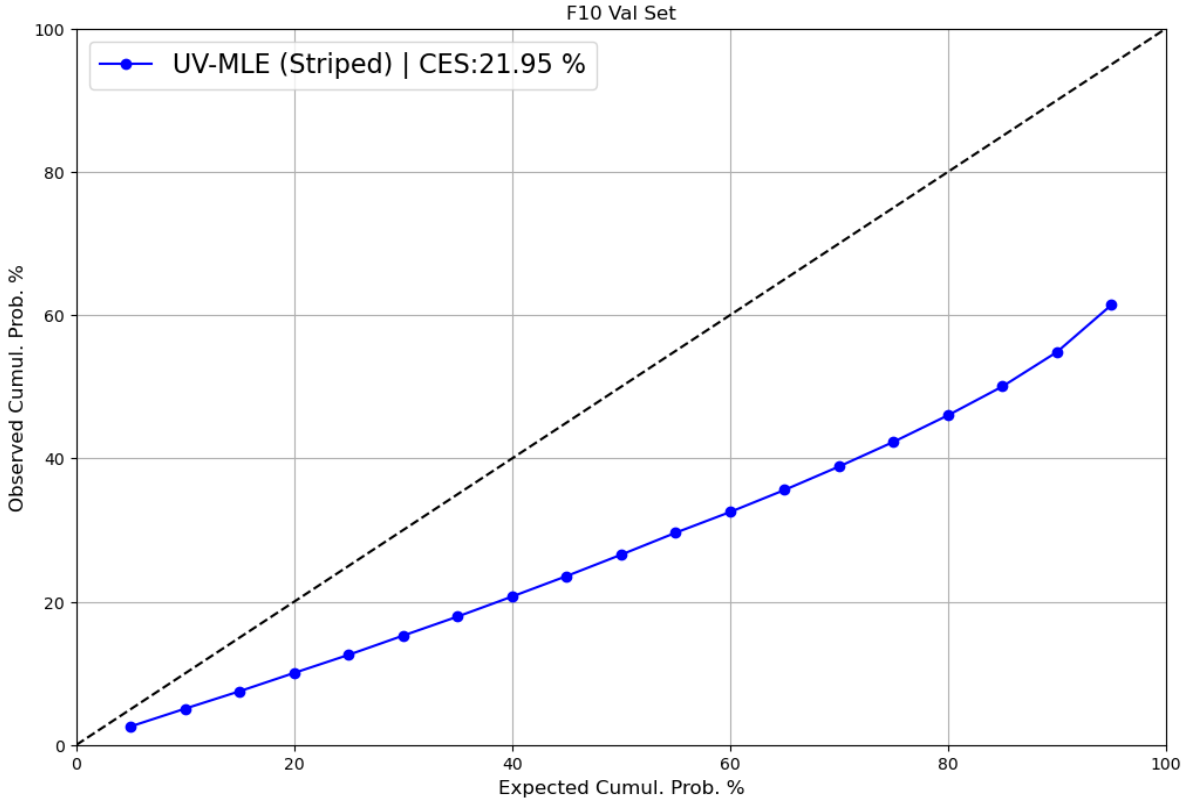
It should be noted that this analysis was done using UV-MLE LSTM which was not the best calibrated model during initial studies. Further study into the performance of striped sampled data on a mixed prediction (such as Dynamic MLP and Multi-Step LSTM) should provide further support for this method. As a result, these calibration curves (and associated CES metric) are not great and should not be considered for operations directly. In general, we see an improvement in training, and similar performance for testing and validation sets. These results are shown WITHOUT any sigma scaling (i.e. these are Raw Outputs). We do also see that the sets are similar, indicating that the model is being tested similarly. Perhaps sigma scaling should be checked as well?

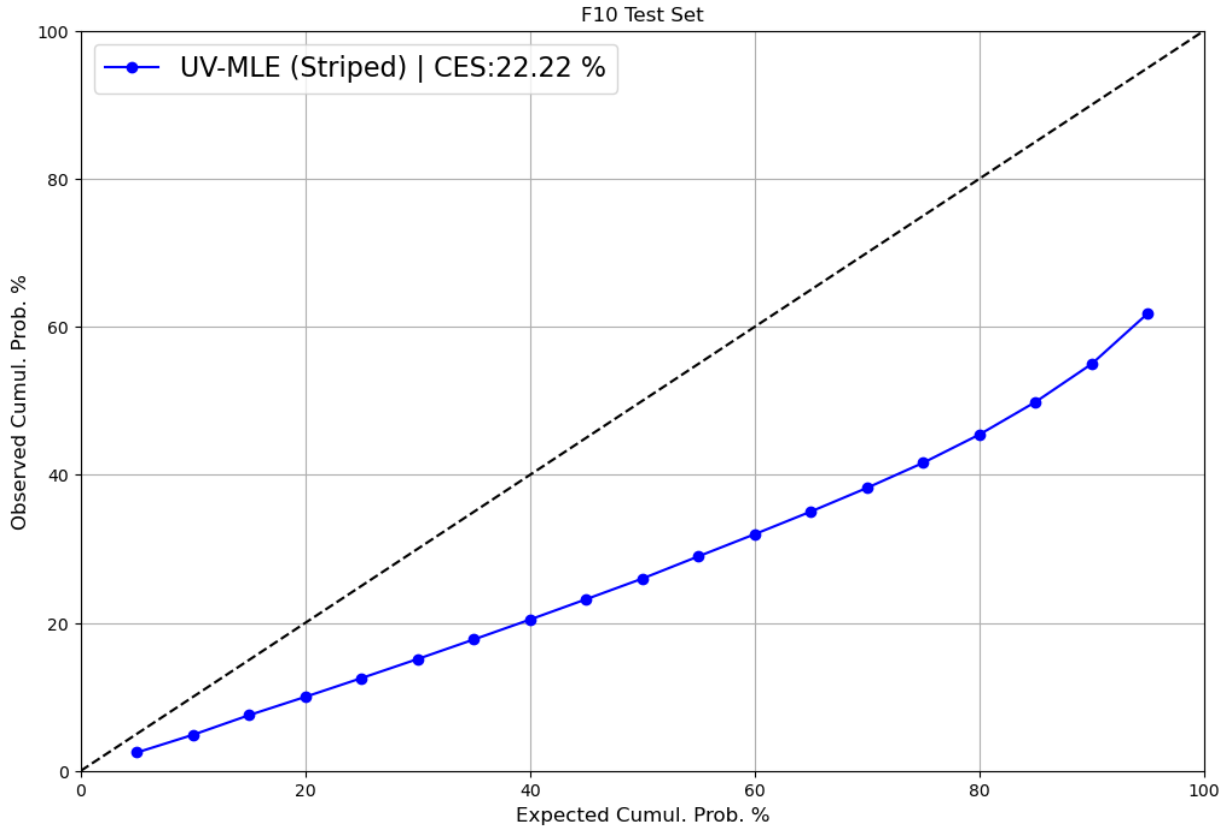
This table shows how holdout and striped sampling CES metrics stack up.

Set	Holdout CES [%]	Striped Sampling CES [%]
Training	38	21.7
Validation	15.5	21.9
Testing	19.4	22.2

Striped Sampling Calibration Curves




F10 Val Set
UV-MLE (Striped) | CES:21.95 %
Observed Cumul. Prob. %
Expected Cumul. Prob. %
0
20
40
60
80
100


F10 Test Set
UV-MLE (Striped) | CES:22.22 %
Observed Cumul. Prob. %
Expected Cumul. Prob. %
0
20
40
60
80
100