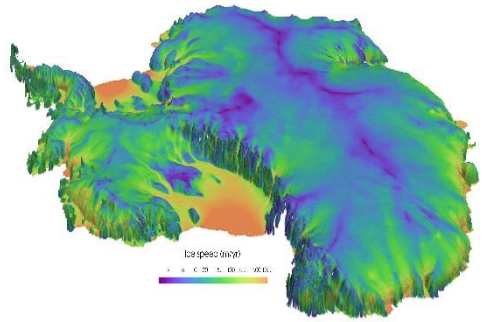
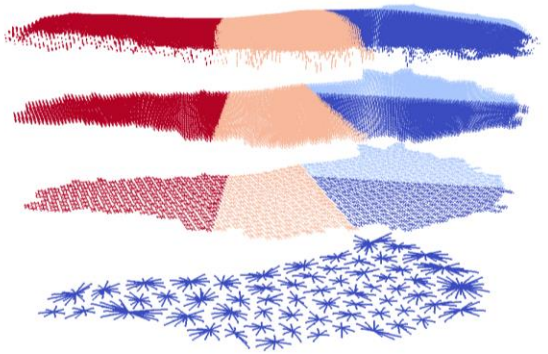
Performance Variation of the Albany Land-Ice Code at Sandia

Kyle Shan, ICME Xplore – Winter 2020

Background

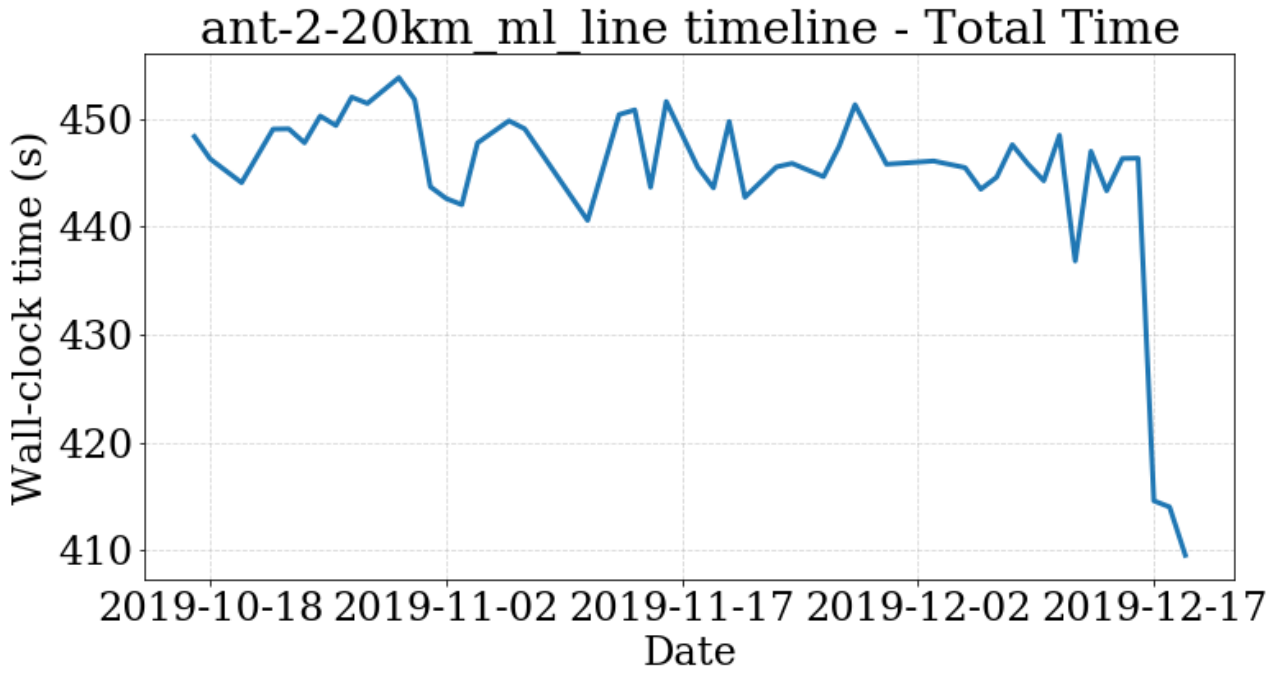
- Land ice modeling is a critical component of Earth System Models, which predict changes in sea level over decades/centuries
- Albany Land-Ice code being developed to run on various HPC architectures
- Ongoing improvements to various code aspects – changes in performance may occur (either intended or unintended)
- **Desired solution: automated review of nightly test cases to identify performance changes**





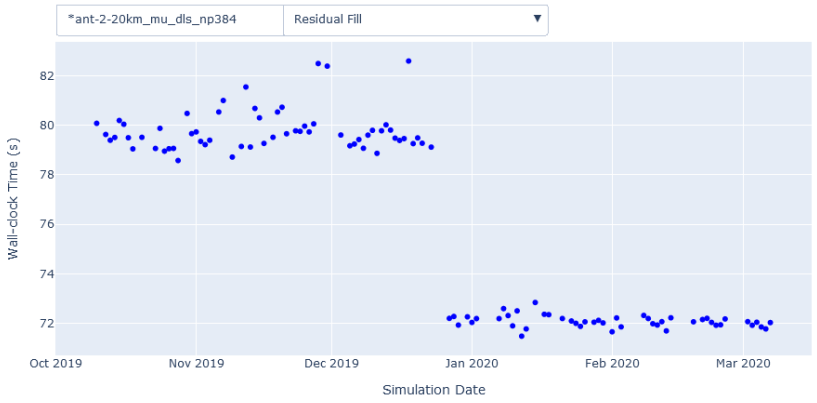
Data

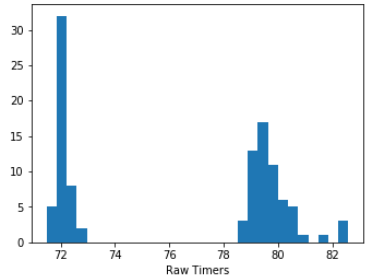
- Nightly test runs of current code base with Antarctic and Greenland ice sheets
- Test cases on Blake (CPU nodes) and Waterman (CPU+GPU nodes) architectures
- Timers measure individual components of the model (e.g. residual/Jacobian fill, preconditioner construction, solver)
- Metadata: Albany/Trilinos repository commit ID, compiler

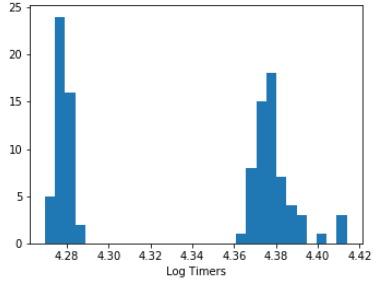


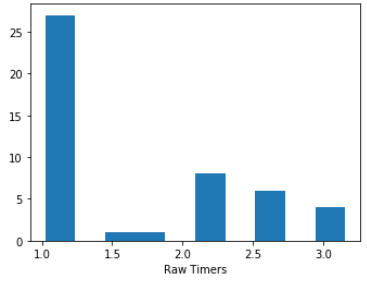
Distribution Assumptions

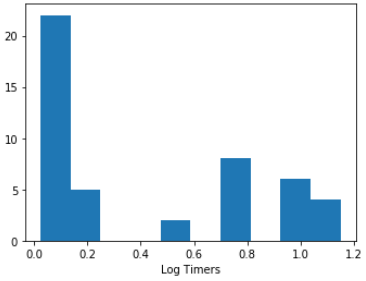
- Timer distribution varies; assume nice distribution mostly to facilitate modeling
- In many cases, lognormal distribution fits slightly better than normal
- Frequent outliers in some datasets





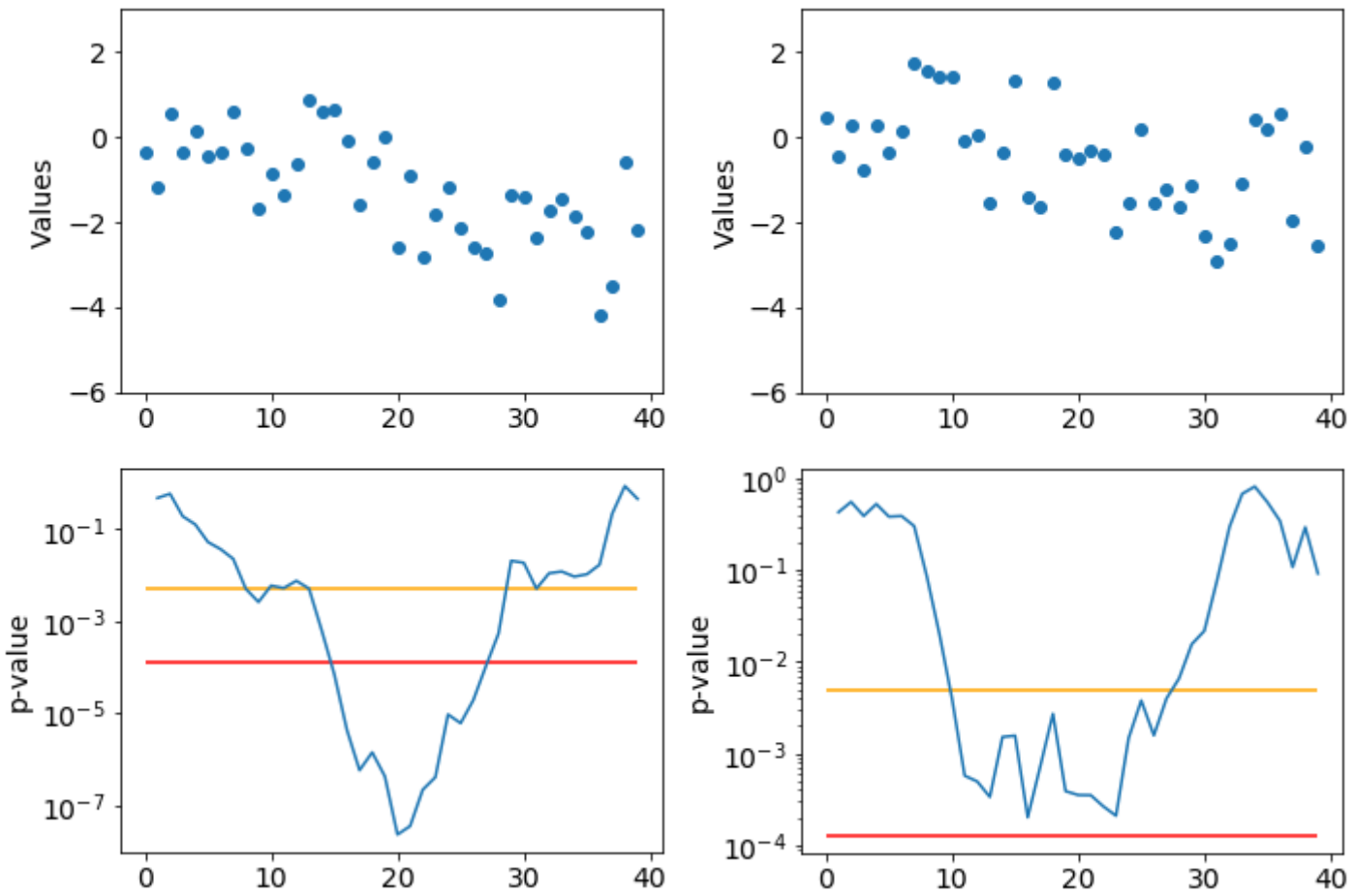






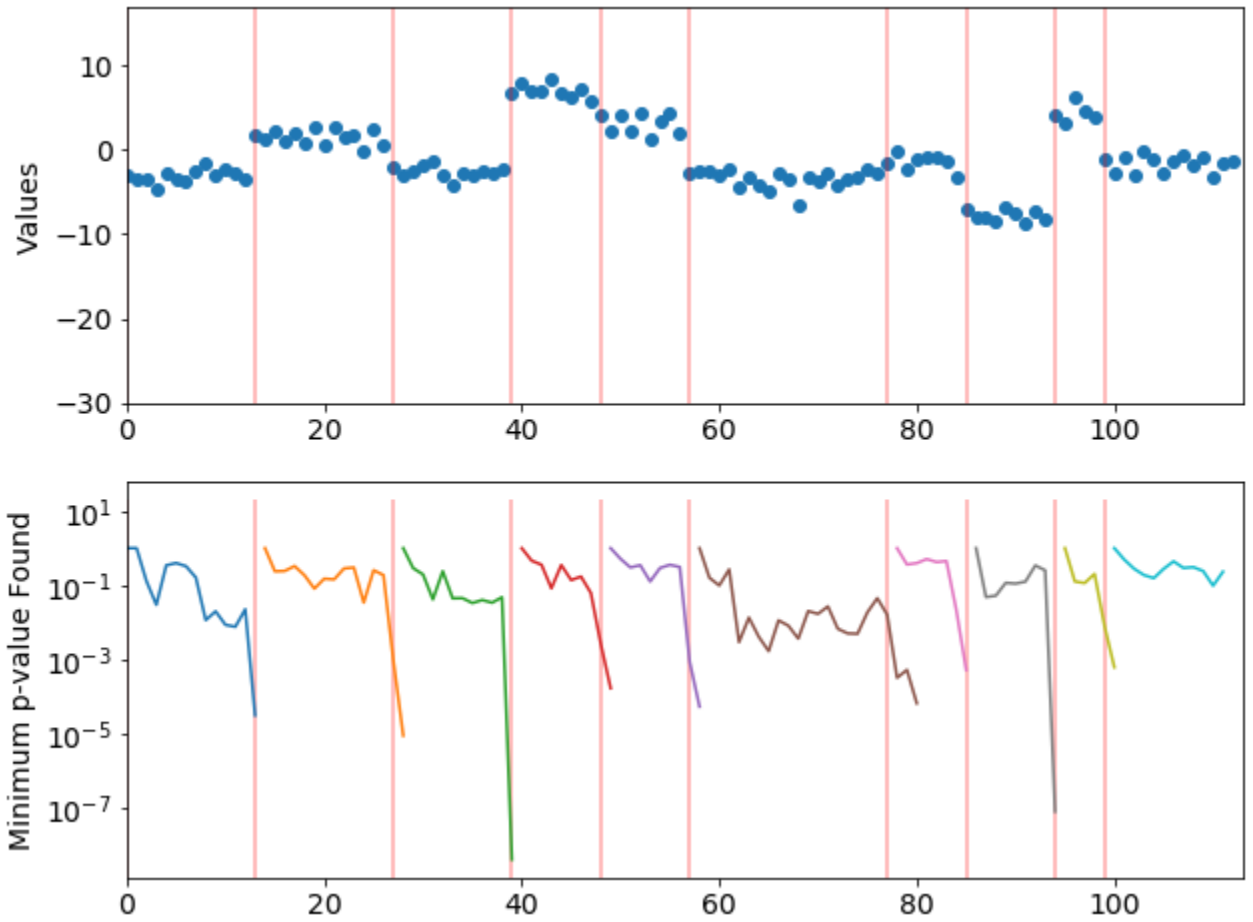
Changepoint Detection

- Given a time series $x_1, \ldots, x_n$, where $x_1, \ldots, x_{\nu-1}$ have distribution f_0 and $x_\nu, \ldots, x_n$ have distribution f_1, can we identify ν?
- Likelihood Ratio idea: changepoint model must explain the data much better than null model
- For normal data, t-test equivalent to likelihood, but must account for multiple hypothesis testing
- Can limit number of tests by only considering k largest changes



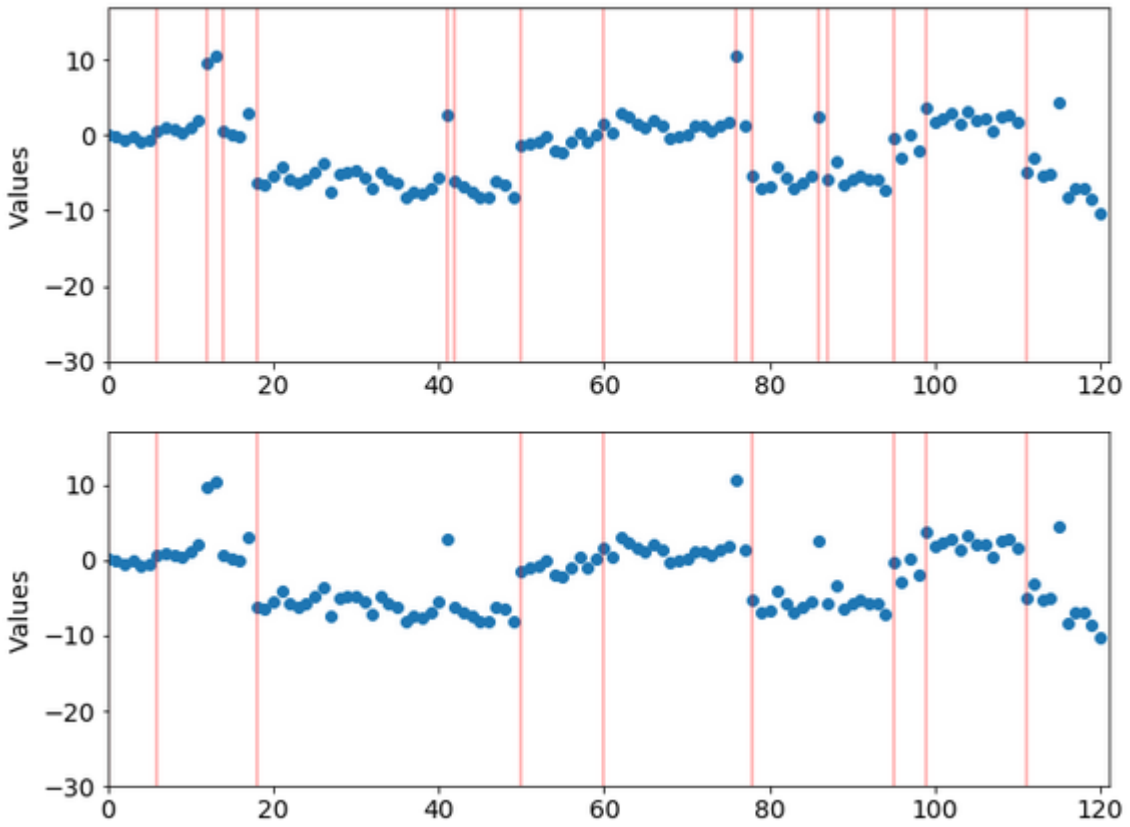
Multiple Changepoints

- Dataset contains an unknown number of changepoints
- Simple extension – check for significant changepoint at each new data point
- Avoids retroactive revisions, but false negatives become highly detrimental



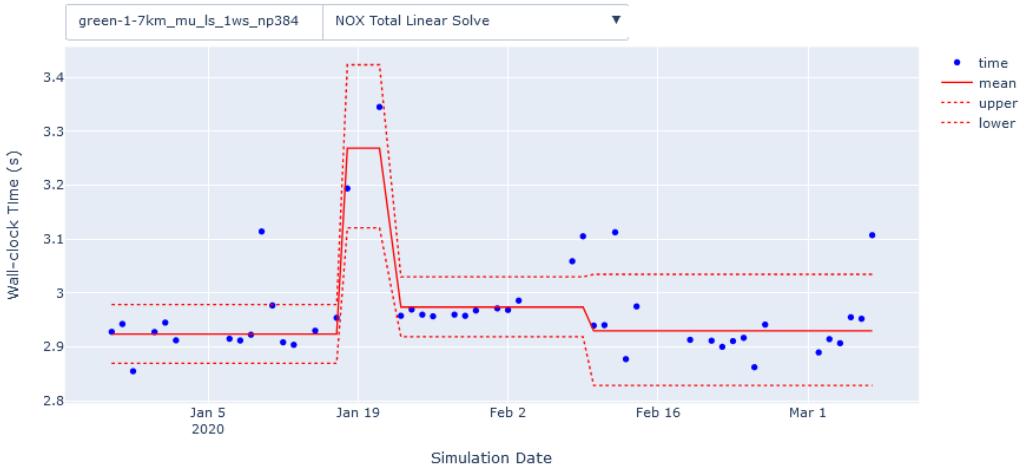
Changepoint or Outlier?

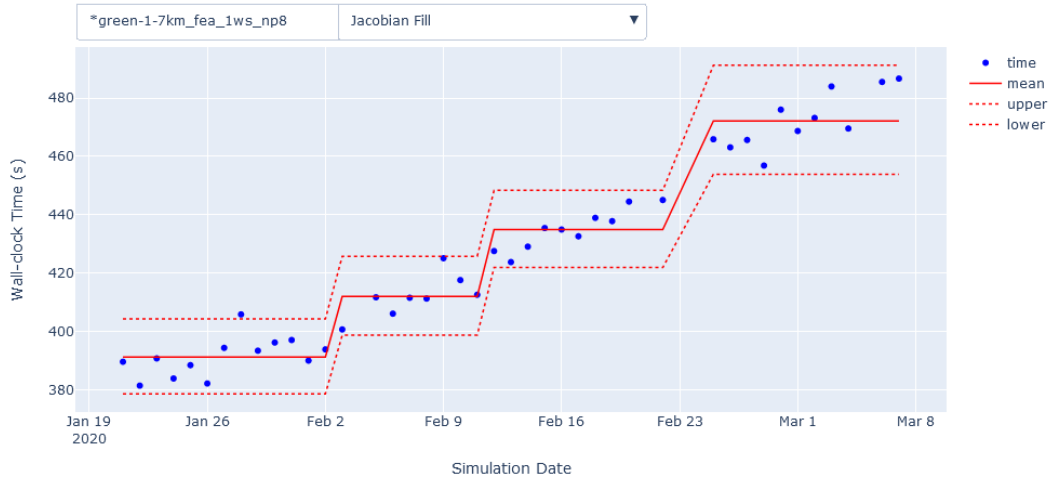
- Outliers *are* changepoints, just not the ones we're interested in; try to lower their influence without reducing power
- Before each t-test, remove potential outliers from each side
- Require consecutive detections of the same changepoint
- Large sample size makes small changes significant; limit lookback window



Subjective Results

- Few instances of known changepoints, so model evaluation is mostly subjective
- Outliers not always ignored successfully but overall, model output seems reasonable



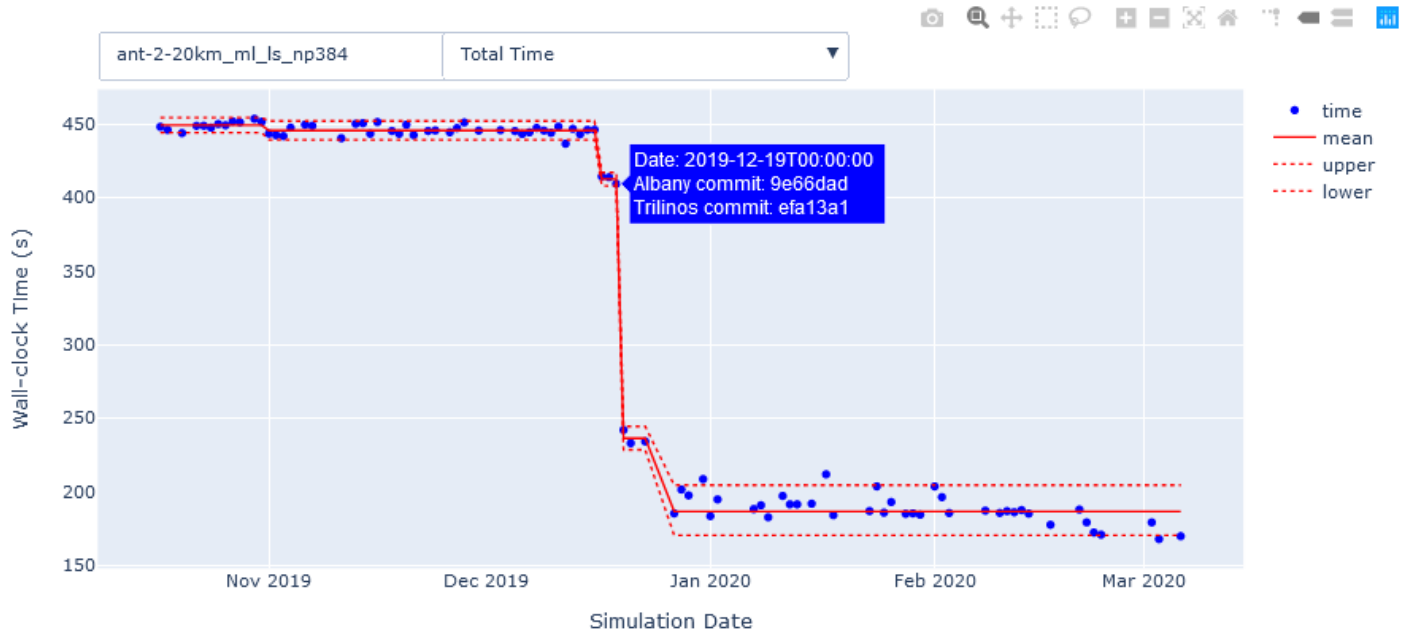


Interactive Monitoring Notebook

- Automatically pulls data from json files
- Computes changepoints and shows recent findings
- Plot of changepoints also provides metadata as hover text
- Interactive drop-downs condense everything into one plot

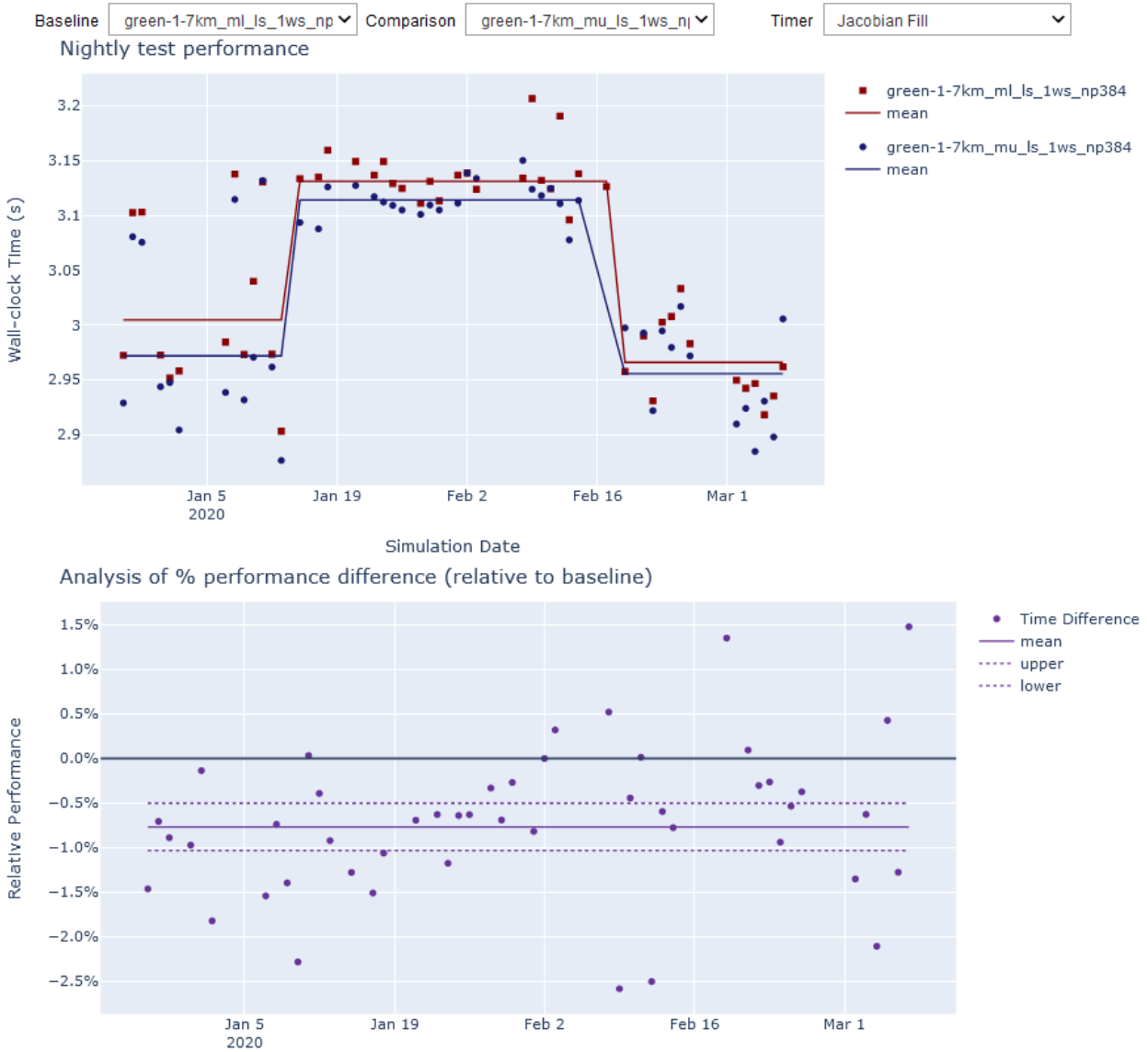
Performance Timelines

```
Events in the most recent 10 days:
02/26/2020:
    ant-2-20km_mu_dls_np384: Total Time
                             NOX Total Preconditioner Construction
                             NOX Total Linear Solve
    green-1-7km_fea_1ws_np8: Residual Fill Evaluate
02/27/2020:
    green-1-7km_fea_mem_np8: Residual Fill
02/29/2020:
    green-1-7km_fea_1ws_np8: Jacobian Fill Export
    green-1-7km_fea_mem_np8: Total Time
                             Total Fill Time
                             Jacobian Fill
                             NOX Total Linear Solve
03/03/2020:
    ant-2-20km_mu_dls_np384: Jacobian Fill
```

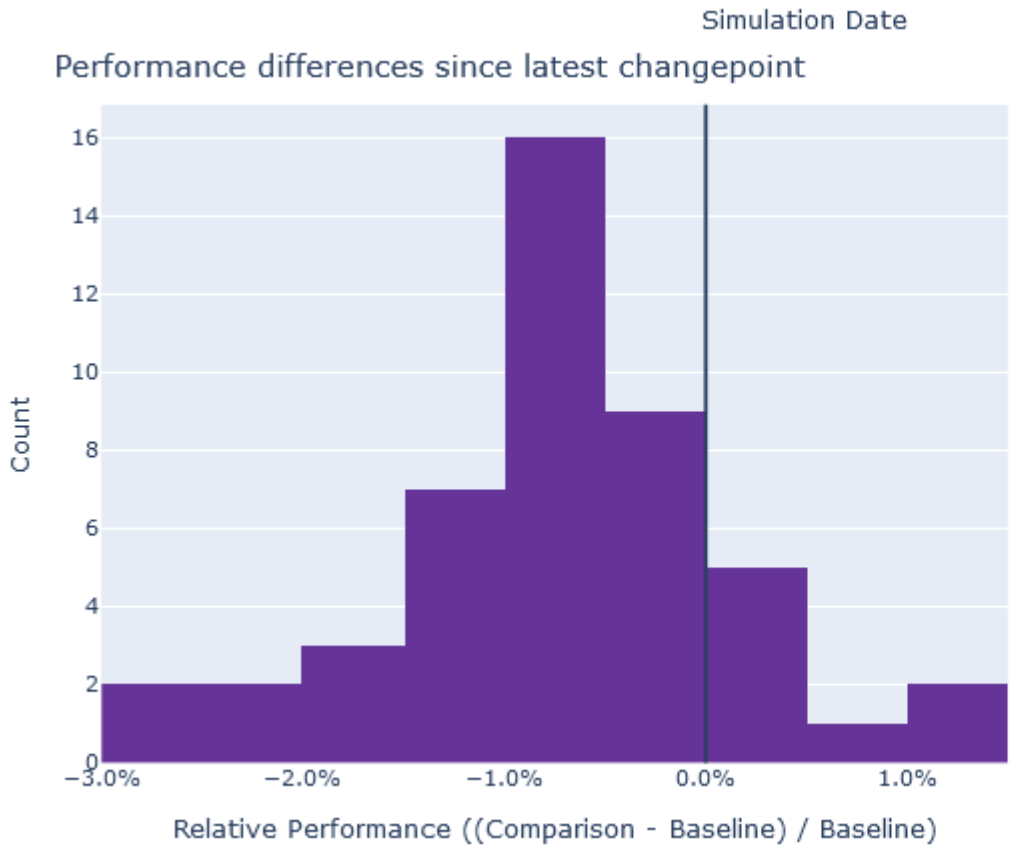


Interactive Comparison Notebook

- Show two test cases side-by-side
- Paired analysis of timers often removes effects of cluster idiosyncrasies, and changes that affect both test cases
- Interactive plots and textbox show relevant statistical findings
- Requires Jupyter notebook for python backend



Interactive Comparison Notebook

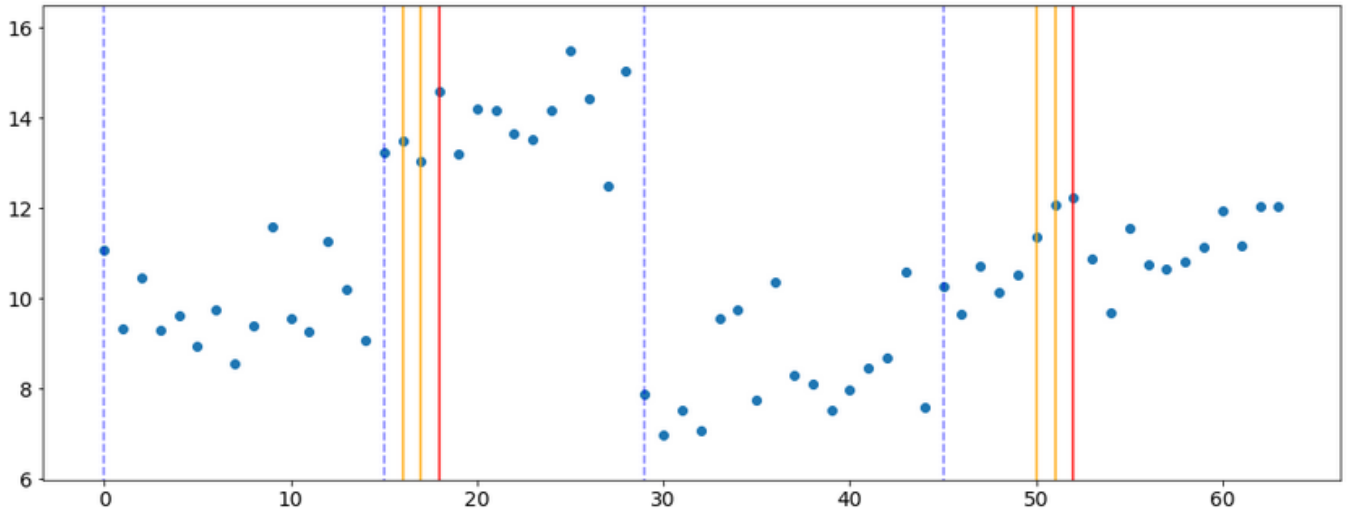



```
Data since latest changepoints
Baseline (since Feb 19):
  green-1-7km_ml_ls_1ws_np384
  N    : 13
  mean: 2.97
  std  : 0.03
Comparison (since Feb 19):
  green-1-7km_mu_ls_1ws_np384
  N    : 13
  mean: 2.96
  std  : 0.04
Paired observations (since Dec 28):
  N    : 47
  mean: -0.81%
  std  : 0.72%
  SE   : 0.10%
  t-test p-value: 1.4e-09***
  99% CI: (-1.03%, -0.50%)
```

Email Report

- Use changepoint detection model in daily monitoring emails to developers
- Warn if significant changepoint detected, and direction of change is positive (tests taking longer)
- Fail if same changepoint has been detected for three consecutive days
- Report mean and standard deviation only since latest changepoint

green-1-7km_fea_1ws_np8 Timers (s)			
Timer	Measured	Mean	Std
Albany Total Time:	636.799	570.683	32.1665
Albany: Setup Time:	42.9695	42.6434	0.979941
Albany: Total Fill Time:	581.76	516.258	31.8427
Albany Fill: Residual:	97.8101	93.4623	2.25448
Albany Residual Fill: Evaluate:	88.7702	86.4673	1.17052
Albany Residual Fill: Export:	2.21035	1.61202	0.30032



Potential Improvements

- Model improvements
  - Instead of reducing sensitivity to outliers, allow them to be detected, but identify whether we revert to the previous distribution afterward
  - Experiment with replacing t-test, or other multiple changepoint detection algorithms
- Notebook improvements
  - Single-timeseries notebook can be viewed as offline HTML, but comparison cannot
  - Web-hosted python app allows maximum portability, but at a cost