

# Clustering Algorithm

update

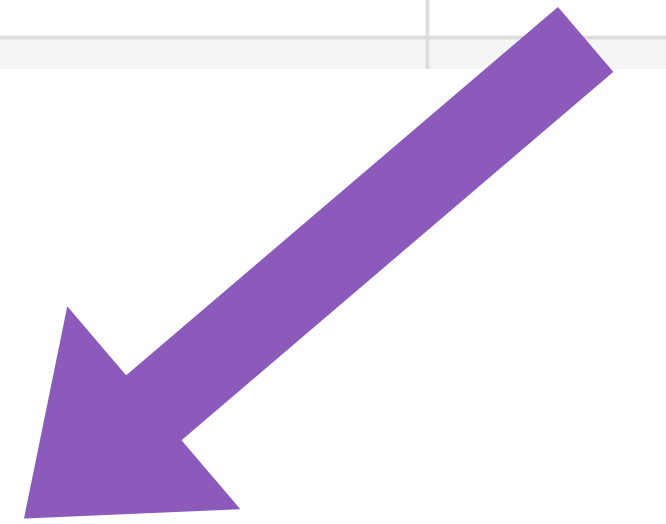
5.26.22

# Scan List from Actual Observation



# “Raw” Data

Longitude	Latitude	Target Name	Min Frequency	Max Frequency	Scan Intent	Polarizations	Temporal Res	Scan Duration
12h36m49.400s	62°11'10.000"	GOODS-N-5	3.9760000 GHz	7.8960000 GHz	["OBSERVE_TARGET"]	["RR, RL, LR, LL"]	1.004	294.2 sec
12h37m2.770s	62°13'52.000"	GOODS-N-7	3.9760000 GHz	7.8960000 GHz	["OBSERVE_TARGET"]	["RR, RL, LR, LL"]	1.003	361 sec
12h36m36.000s	62°13'52.000"	GOODS-N-3	3.9760000 GHz	7.8960000 GHz	["OBSERVE_TARGET"]	["RR, RL, LR, LL"]	1.003	361 sec
13h31m8.288s	30°30'32.959"	J1331+3030	3.9760000 GHz	7.8960000 GHz	["CALIBRATE_BANDPASS","CALIBRATE_FLUX"]	["RR, RL, LR, LL"]	1.002	598.35 sec
12h36m49.400s	62°14'46.000"	GOODS-N-2	3.9760000 GHz	7.8960000 GHz	["OBSERVE_TARGET"]	["RR, RL, LR, LL"]	1.003	361.05 sec
12h36m49.400s	62°11'10.000"	GOODS-N-5	3.9760000 GHz	7.8960000 GHz	["OBSERVE_TARGET"]	["RR, RL, LR, LL"]	1.003	361.05 sec
12h17m11.019s	58°35'26.248"	J1217+5835	3.9760000 GHz	7.8960000 GHz	["SYSTEM_CONFIGURATION"]	["RR, RL, LR, LL"]	1.014	59.85 sec
12h17m11.019s	58°35'26.248"	J1217+5835	4.8320000 GHz	4.9600000 GHz	["SYSTEM_CONFIGURATION"]	["RR, RL, LR, LL"]	1.003	508.6 sec

























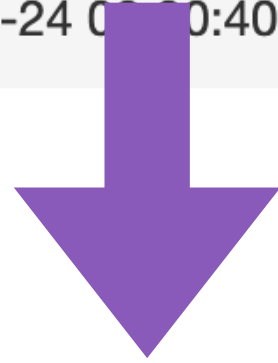
Requested Time = Scan Duration for Scan Intents OBSERVE\_TARGET

# Scan List from Actual Observation



# “Raw” Data

↕ Archive File	↕ Project	↕ Instrument	↕ Observation Start	↕ Observation Stop	↕ File Size	Array Config	Bands	Type	Cals	Scans
 19A-401.sb37874239.eb37903476.58888.37139646991	19A-401	EVLA	2020-02-09 08:54:49	2020-02-09 10:54:12	146.881 GB	C	C	visibility	 1	29
 19A-401.sb36960865.eb37113335.58720.76762262732	19A-401	EVLA	2019-08-25 18:25:23	2019-08-25 22:24:40	330.121 GB	A	C	visibility	 1	60
 19A-401.sb36985043.eb37111985.58719.81641645833	19A-401	EVLA	2019-08-24 19:35:39	2019-08-24 23:34:55	330.143 GB	A	C	visibility	 1	60
 19A-401.sb36985373.eb37107930.58718.799060578705	19A-401	EVLA	2019-08-23 19:10:39	2019-08-23 23:09:55	306.184 GB	A	C	visibility	 1	60
 19A-401.sb36960700.eb37083390.58713.834670428245	19A-401	EVLA	2019-08-18 20:01:56	2019-08-19 00:01:12	306.184 GB	A	C	visibility	 1	60
 19A-401.sb36984878.eb37049234.58705.81565862268	19A-401	EVLA	2019-08-10 19:34:33	2019-08-10 23:33:53	330.219 GB	A	C	visibility	 1	60
 19A-401.sb36985208.eb37045620.58704.85853885417	19A-401	EVLA	2019-08-09 20:36:18	2019-08-10 00:35:37	330.194 GB	A	C	visibility	 1	60
 19A-401.sb36960284.eb37029385.58701.80962503472	19A-401	EVLA	2019-08-06 19:25:52	2019-08-06 23:24:39	330.124 GB	A	C	visibility	 1	60
 19A-401.sb36960535.eb37010071.58699.87413458333	19A-401	EVLA	2019-08-04 20:58:45	2019-08-05 01:28:01	345.944 GB	A	C	visibility	 1	69
 19A-401.sb36613321.eb36616533.58603.135411944444	19A-401	EVLA	2019-04-30 03:15:00	2019-04-30 07:14:16	330.143 GB	B	C	visibility	 1	60
 19A-401.sb36305804.eb36602150.58597.097676840276	19A-401	EVLA	2019-04-24 00:00:40	2019-04-24 06:19:59	330.148 GB	B	C	visibility	 1	60



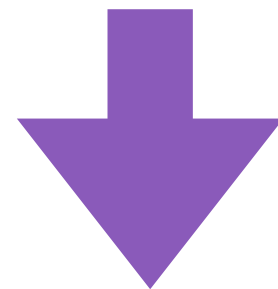
All Execution Blocks (EB) per proposal generate Science Target List

# “Raw” Data

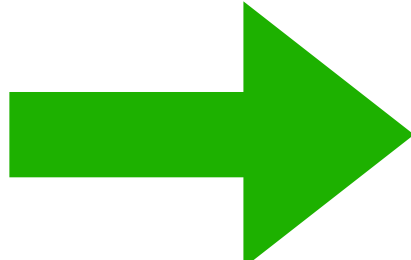
Archive File	Project	Instrument	Observation Start	Observation Stop	File Size	Array Config	Bands	Type	Cals	Scans
19A-401.sb37874239.eb37903476.58888.37139646991	19A-401	EVLA	2020-02-09 08:54:49	2020-02-09 10:54:12	146.881 GB	C	C	visibility	1	29
19A-401.sb36960865.eb37113335.58720.76762262732	19A-401	EVLA	2019-08-25 18:25:23	2019-08-25 22:24:40	330.121 GB	A	C	visibility	1	60
19A-401.sb36985043.eb37111985.58719.81641645833	19A-401	EVLA	2019-08-24 19:35:39	2019-08-24 23:34:55	330.143 GB	A	C	visibility	1	60
19A-401.sb36985373.eb37107930.58718.799060578705	19A-401	EVLA	2019-08-23 19:10:39	2019-08-23 23:09:55	306.184 GB	A	C	visibility	1	60
19A-401.sb36960700.eb37083390.58713.834670428245	19A-401	EVLA	2019-08-18 20:01:56	2019-08-19 00:01:12	306.184 GB	A	C	visibility	1	60
19A-401.sb36984878.eb37049234.58705.81565862268	19A-401	EVLA	2019-08-10 19:34:33	2019-08-10 23:33:53	330.219 GB	A	C	visibility	1	60
19A-401.sb36985208.eb37045620.58704.85853885417	19A-401	EVLA	2019-08-09 20:36:18	2019-08-10 00:35:37	330.194 GB	A	C	visibility	1	60
19A-401.sb36960284.eb37029385.58701.80962503472	19A-401	EVLA	2019-08-06 19:25:52	2019-08-06 23:24:39	330.124 GB	A	C	visibility	1	60
19A-401.sb36960535.eb37010071.58699.87413458333	19A-401	EVLA	2019-08-04 20:58:45	2019-08-05 01:28:01	345.944 GB	A	C	visibility	1	69
19A-401.sb36613321.eb36616533.58603.135411944444	19A-401	EVLA	2019-04-30 03:15:00	2019-04-30 07:14:16	330.143 GB	B	C	visibility	1	60
19A-401.sb36305804.eb36602150.58597.097676840276	19A-401	EVLA	2019-04-24 02:20:40	2019-04-24 06:19:59	330.148 GB	B	C	visibility	1	60

Python Script to Generate Science Targets

(Mock) ObservingStrategy



- Assign clusters using Execution Block as template
- Use flat calibration overhead factor to estimate duration



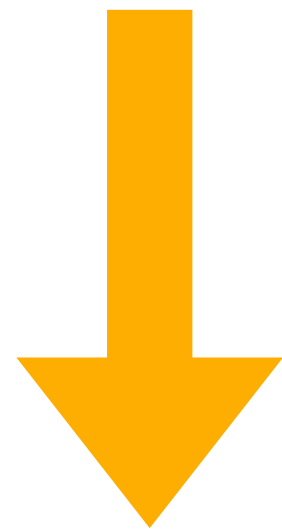
# “Real” Data

Cluster ID	ScienceTarget ID	Source Name	Total Requested Time (hr)	Req. Time per Repeat Count (hr)	Repeat Count	Total Duration (hr)	Duration per RC (hr)	Cluster Size	Band	Facility/Config
0	[1, 4, 7, 10, 13, 16, 19, 22]	['GOODS-N-1', 'GOODS-N-2', 'GOODS-N-3', 'GOODS-N-4', 'GOODS-N-5', 'GOODS-N-6', 'GOODS-N-7', 'J124129+6020']	5.99	3.00	[2]	9.03	4.51	8	C	B-array
1	[0, 3, 6, 9, 12, 15, 18, 21]	['GOODS-N-1', 'GOODS-N-2', 'GOODS-N-3', 'GOODS-N-4', 'GOODS-N-5', 'GOODS-N-6', 'GOODS-N-7', 'J124129+6020']	24.41	3.07	[7, 8]	36.95	4.62	8	C	A-array
2	[2, 5, 8, 11, 14, 17, 20]	['GOODS-N-1', 'GOODS-N-2', 'GOODS-N-3', 'GOODS-N-4', 'GOODS-N-5', 'GOODS-N-6', 'GOODS-N-7']	1.27	1.27	[1]	1.97	1.97	7	C	C-array

## Observing Strategy

### Selection of Partition Plan Considers

- How many Frequencies requested
- Composition of requested Frequencies
- Requested Time per Band
- How many Sources
  - Is project a Mosaic



## Science Target List

- Requested Time, Configuration, Frequency, Coordinate
- Partition Plan

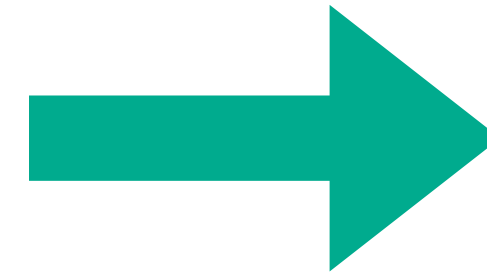
\*Items in blue are examples and are NOT definitive. These examples showcase the flexibility of the algorithm and the breadth of customization possible.

## Partition Plans

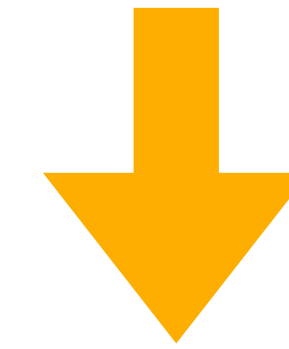
- Configuration Filter
  - e.g., different VLA configurations cannot be clustered together
- Distance Filter
  - e.g., clusters must be < **10 deg** in size
- Frequency Filter
  - No Filter Applied
  - Hierarchical Filters (selected examples)
    - **Q- and Ka-band cannot cluster with any other band**
    - **Let Ka- & K-band cluster if Ku-band isn't requested**
    - **Let C- & S-band cluster if X-band isn't requested**
    - **Let X- & C-band cluster**
- Temporal Filter (selected examples)
  - Clusters must be < **3 hours** (including Overhead)
  - Maximize cluster (temporal) size - make clusters as close to maximum duration as possible within the constraints
  - SetupTime accounts for < **5% of duration**
  - Minimum duration of a sub scan > **30 seconds**
  - Prioritize sub scan durations of ~**20 minutes**, if applicable
  - Sources cannot have RequestedTime > time the source is above the specified elevation

# Science Target List

- No knowledge of how “Real” data is clustered
- Requested Time, Configuration, Frequency, Coordinate Information
- Partition Plan



Clustering Algorithm



“Model” Data

Cluster ID	ScienceTarget ID	Source Name	Total Requested Time (hr)	Req. Time per Repeat Count (hr)	Repeat Count	Total Duration (hr)	Duration per RC (hr)	Cluster Size	Band	Facility/Config
0	[2, 5, 8, 11, 14, 17, 20]	['GOODS-N-1', 'GOODS-N-2', 'GOODS-N-3', 'GOODS-N-4', 'GOODS-N-5', 'GOODS-N-6', 'GOODS-N-7']	1.27	1.27	1	1.97	1.97	7	C	C-array
1	[0, 3, 6, 9, 12, 15, 18]	['GOODS-N-1', 'GOODS-N-2', 'GOODS-N-3', 'GOODS-N-4', 'GOODS-N-5', 'GOODS-N-6', 'GOODS-N-7']	23.45	1.95	12	35.65	2.97	7	C	A-array
2	[21]	['J124129+6020']	0.96	0.96	1	1.41	1.41	1	C	A-array
3	[1, 4, 7, 10, 13, 16, 19]	['GOODS-N-1', 'GOODS-N-2', 'GOODS-N-3', 'GOODS-N-4', 'GOODS-N-5', 'GOODS-N-6', 'GOODS-N-7']	5.72	1.91	3	8.70	2.90	7	C	B-array
4	[22]	['J124129+6020']	0.27	0.27	1	0.40	0.40	1	C	B-array

Cluster ID	ScienceTarget ID	Source Name	Total Requested Time (hr)	Req. Time per Repeat Count (hr)	Repeat Count	Total Duration (hr)	Duration per RC (hr)	Cluster Size	Band	Facility/Config
0	[1, 4, 7, 10, 13, 16, 19, 22]	['GOODS-N-1', 'GOODS-N-2', 'GOODS-N-3', 'GOODS-N-4', 'GOODS-N-5', 'GOODS-N-6', 'GOODS-N-7', 'J124129+6020']	5.99	3.00	[2]	9.03	4.51	8	C	B-array
1	[0, 3, 6, 9, 12, 15, 18, 21]	['GOODS-N-1', 'GOODS-N-2', 'GOODS-N-3', 'GOODS-N-4', 'GOODS-N-5', 'GOODS-N-6', 'GOODS-N-7', 'J124129+6020']	24.41	3.07	[7, 8]	36.95	4.62	8	C	A-array
2	[2, 5, 8, 11, 14, 17, 20]	['GOODS-N-1', 'GOODS-N-2', 'GOODS-N-3', 'GOODS-N-4', 'GOODS-N-5', 'GOODS-N-6', 'GOODS-N-7']	1.27	1.27	[1]	1.97	1.97	7	C	C-array

“Real” Data

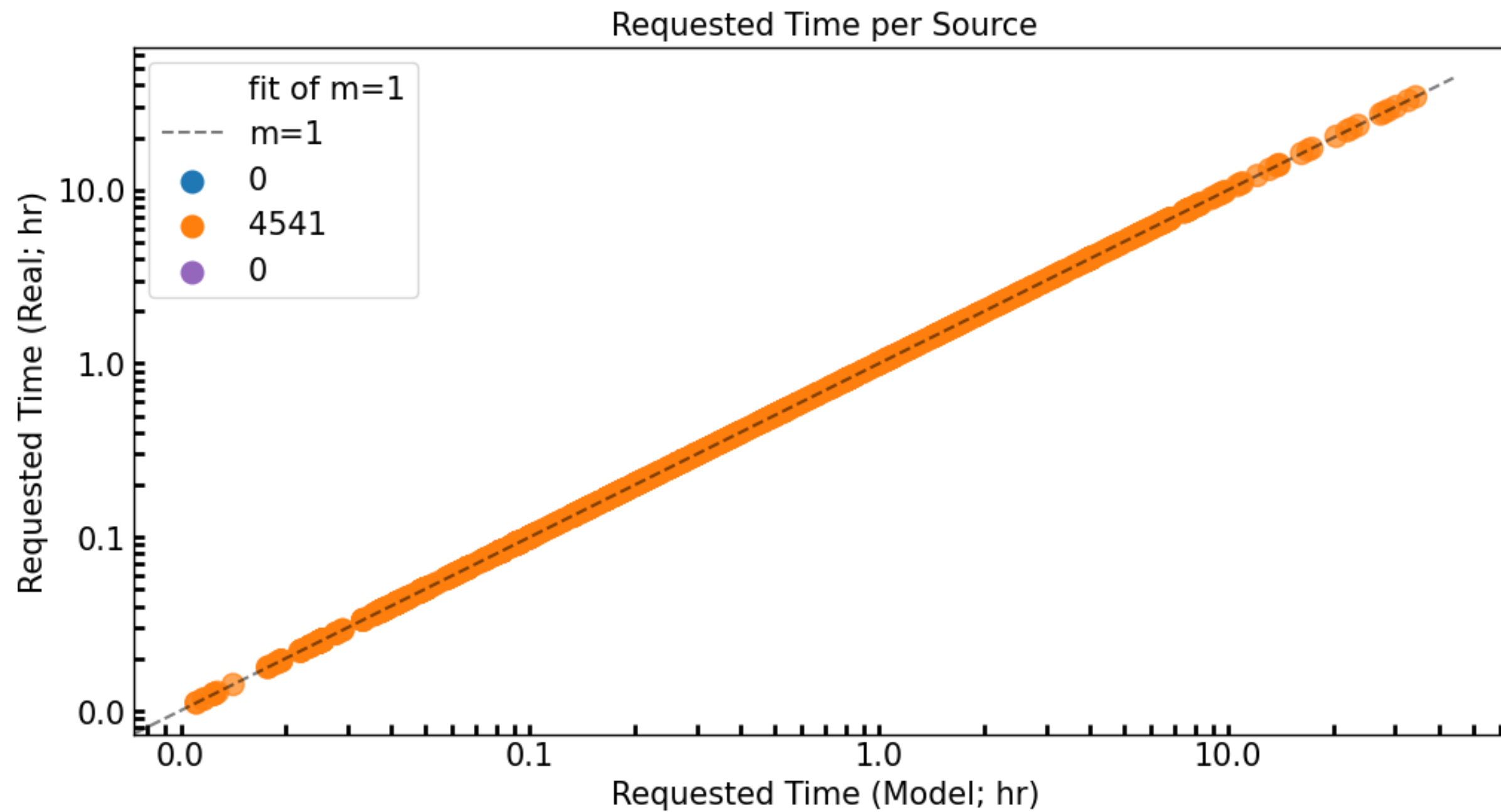
Compare:

- Number of Clusters
- Number of Observation Specifications (or Execution Blocks) =
- Clusters x Repeat Counts
- Total Duration
- Total Requested Time
- Duration of an Observation Specification

Cluster ID	ScienceTarget ID	Source Name	Total Requested Time (hr)	Req. Time per Repeat Count (hr)	Repeat Count	Total Duration (hr)	Duration per RC (hr)	Cluster Size	Band	Facility/Config
0	[2, 5, 8, 11, 14, 17, 20]	['GOODS-N-1', 'GOODS-N-2', 'GOODS-N-3', 'GOODS-N-4', 'GOODS-N-5', 'GOODS-N-6', 'GOODS-N-7']	1.27	1.27	1	1.97	1.97	7	C	C-array
1	[0, 3, 6, 9, 12, 15, 18]	['GOODS-N-1', 'GOODS-N-2', 'GOODS-N-3', 'GOODS-N-4', 'GOODS-N-5', 'GOODS-N-6', 'GOODS-N-7']	23.45	1.95	12	35.65	2.97	7	C	A-array
2	[21]	['J124129+6020']	0.96	0.96	1	1.41	1.41	1	C	A-array
3	[1, 4, 7, 10, 13, 16, 19]	['GOODS-N-1', 'GOODS-N-2', 'GOODS-N-3', 'GOODS-N-4', 'GOODS-N-5', 'GOODS-N-6', 'GOODS-N-7']	5.72	1.91	3	8.70	2.90	7	C	B-array
4	[22]	['J124129+6020']	0.27	0.27	1	0.40	0.40	1	C	B-array

“Model” Data

Does the **model** request the same Requested Time as the **real proposal**? (not total time!)

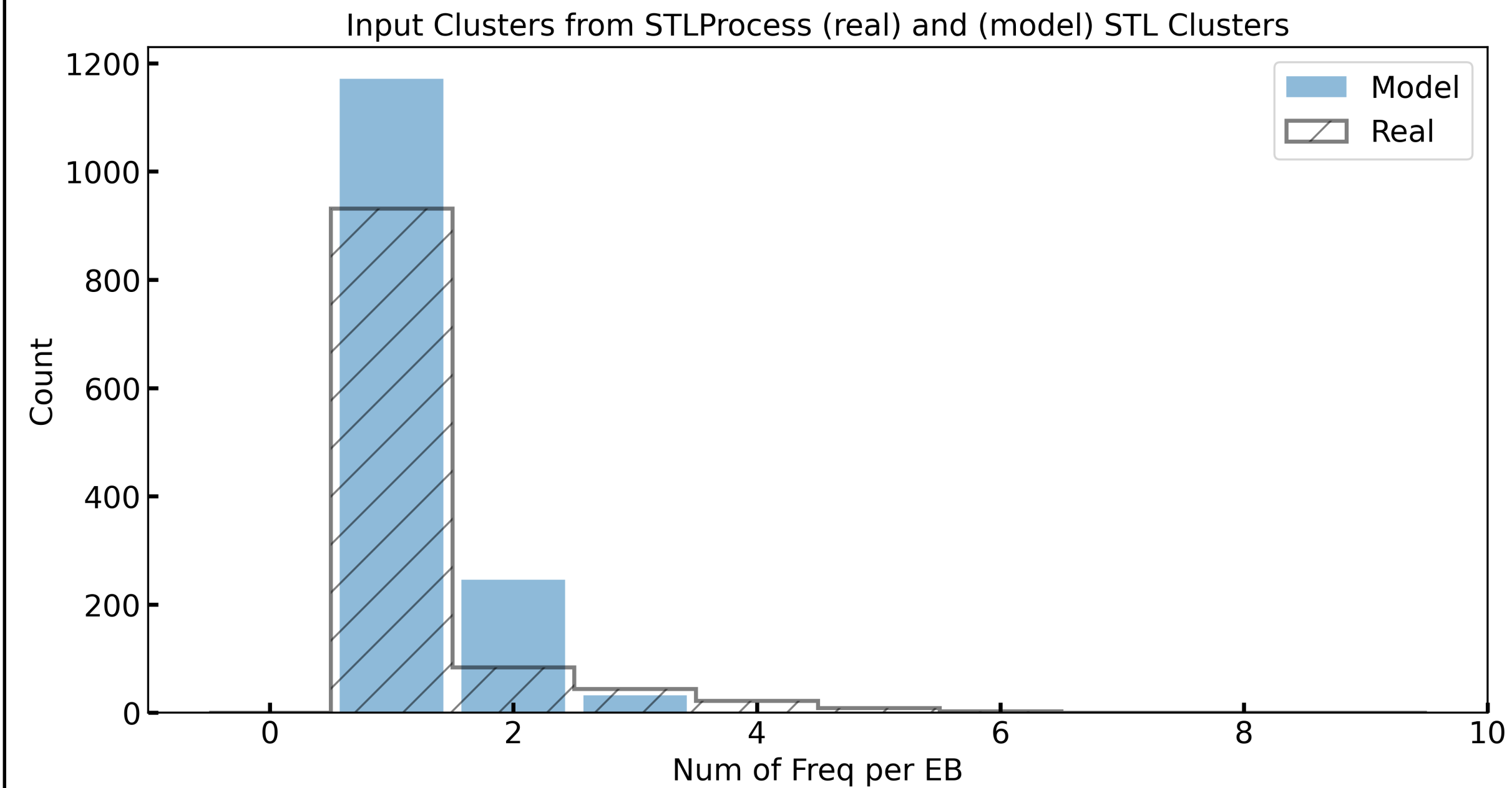


“Real” Data

“Model” Data

yes!

How does the **model** cluster frequency compared to **real proposal**?

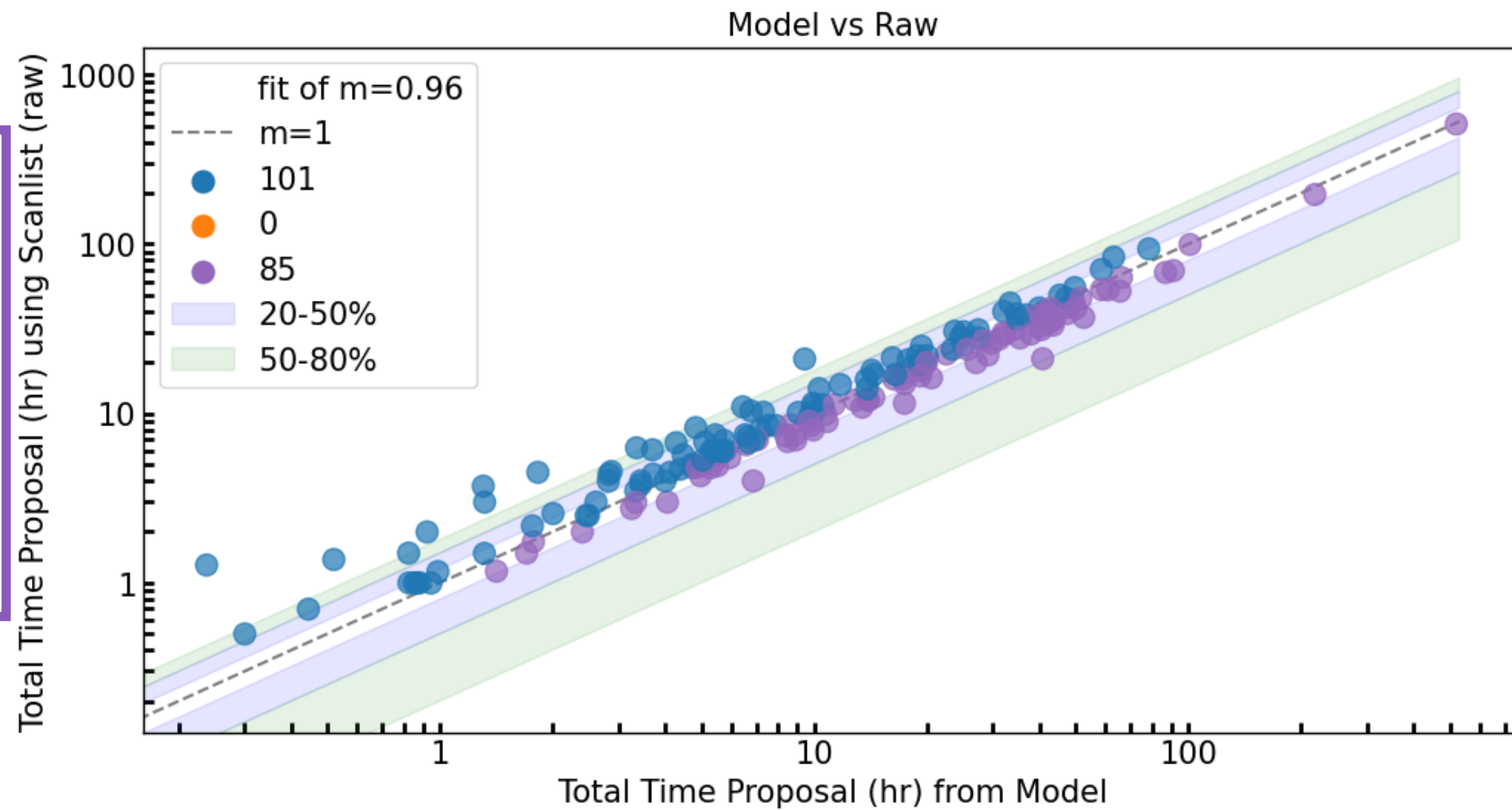


- The model favors  $< 3$  bands
- The model does not put 4+ bands in one cluster
- But there are only a few instances of users putting 4+ bands in a single cluster

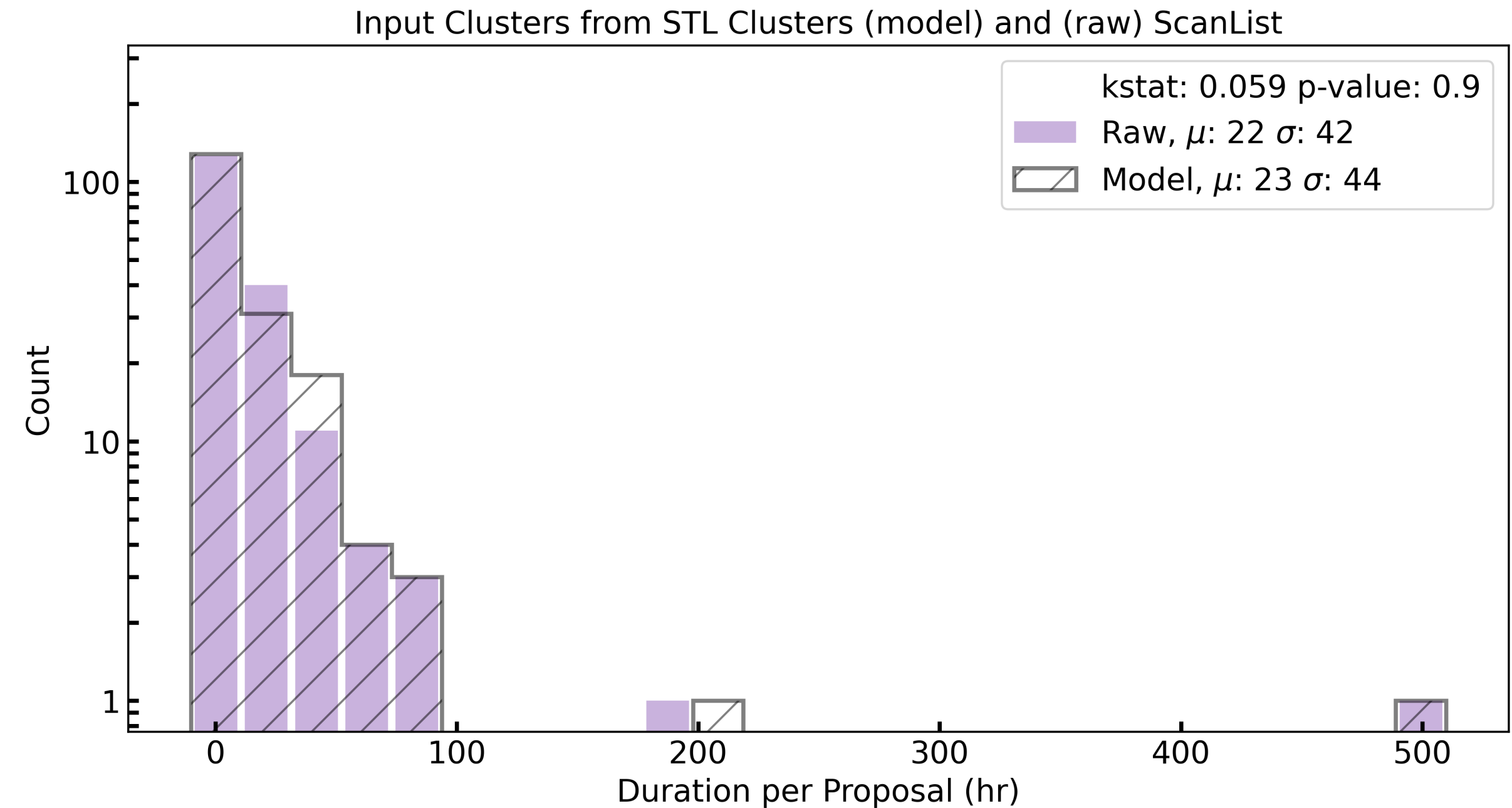


# How much Total Time does the **model** predict compared to the **raw Scan List**?

“Raw” Data



“Model” Data



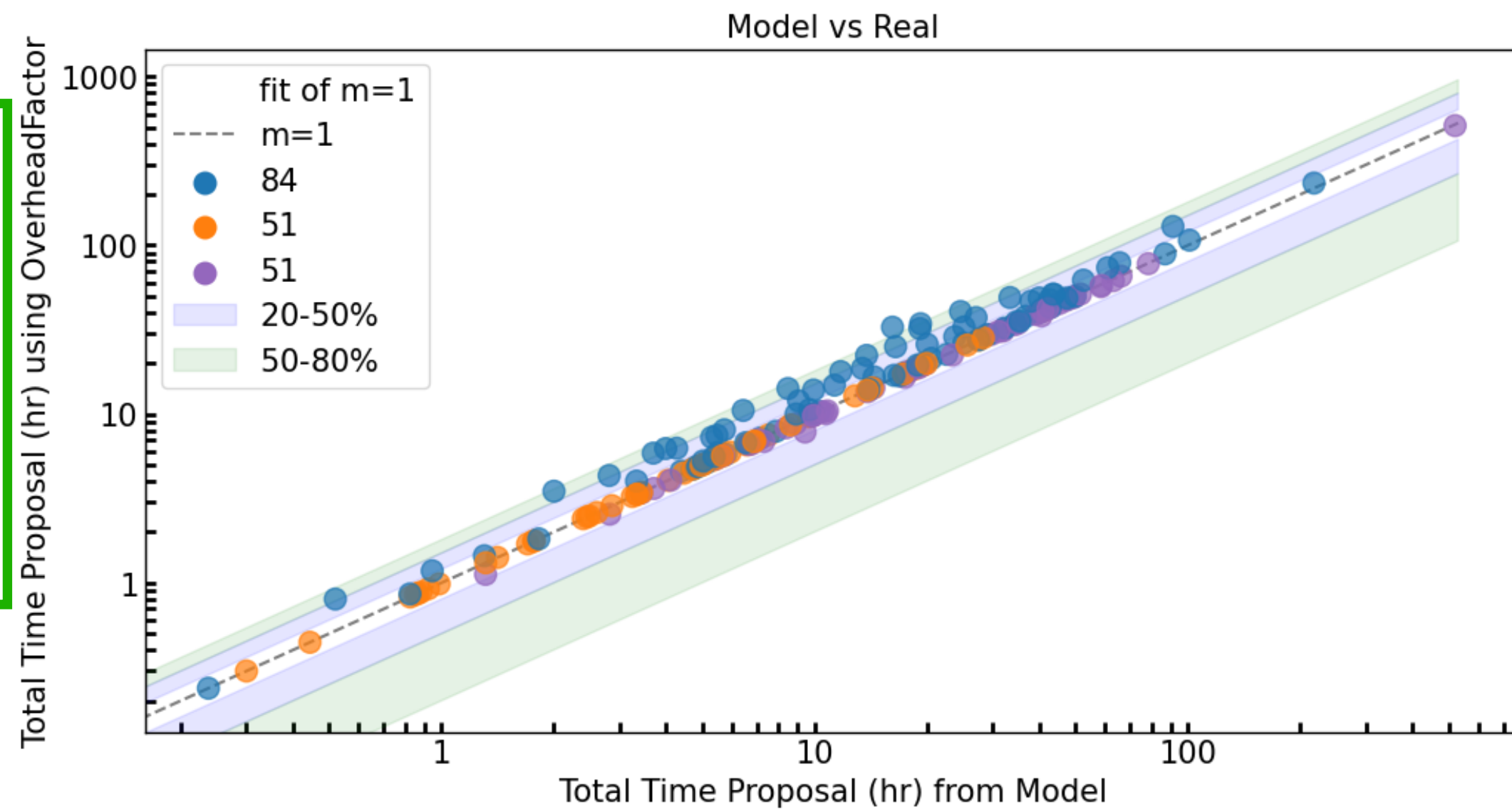
~~“Model” Data~~

“Raw” Data

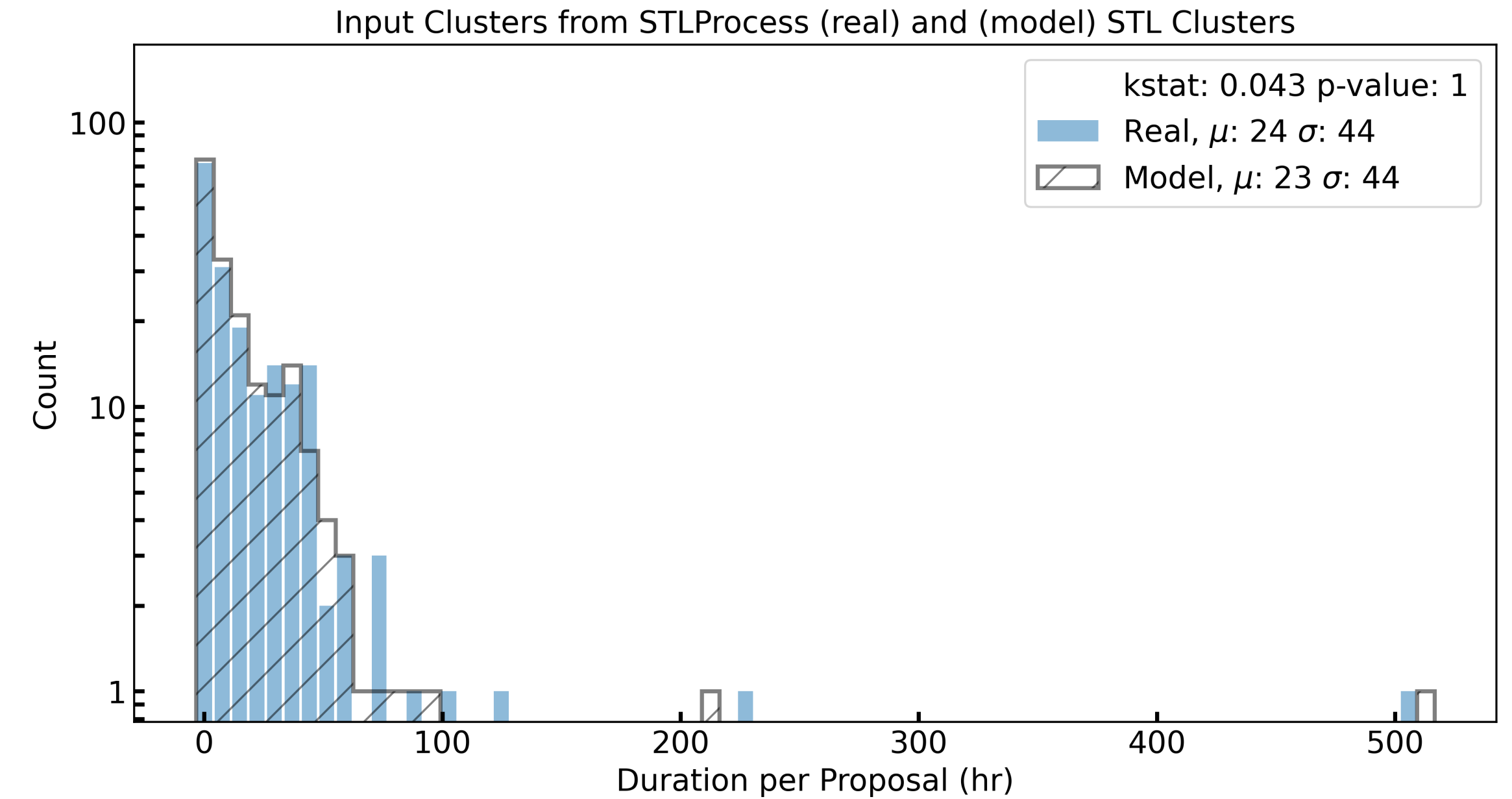
- The model tends to predict more Time per proposal than the Raw Scan List (i.e., what actually ran on the antenna).

How much Total Time does the **model** give to a proposal compared to the **real proposal**, when using the same **Calibration overhead factor**?

“Real” Data



“Model” Data



~~“Model” Data~~

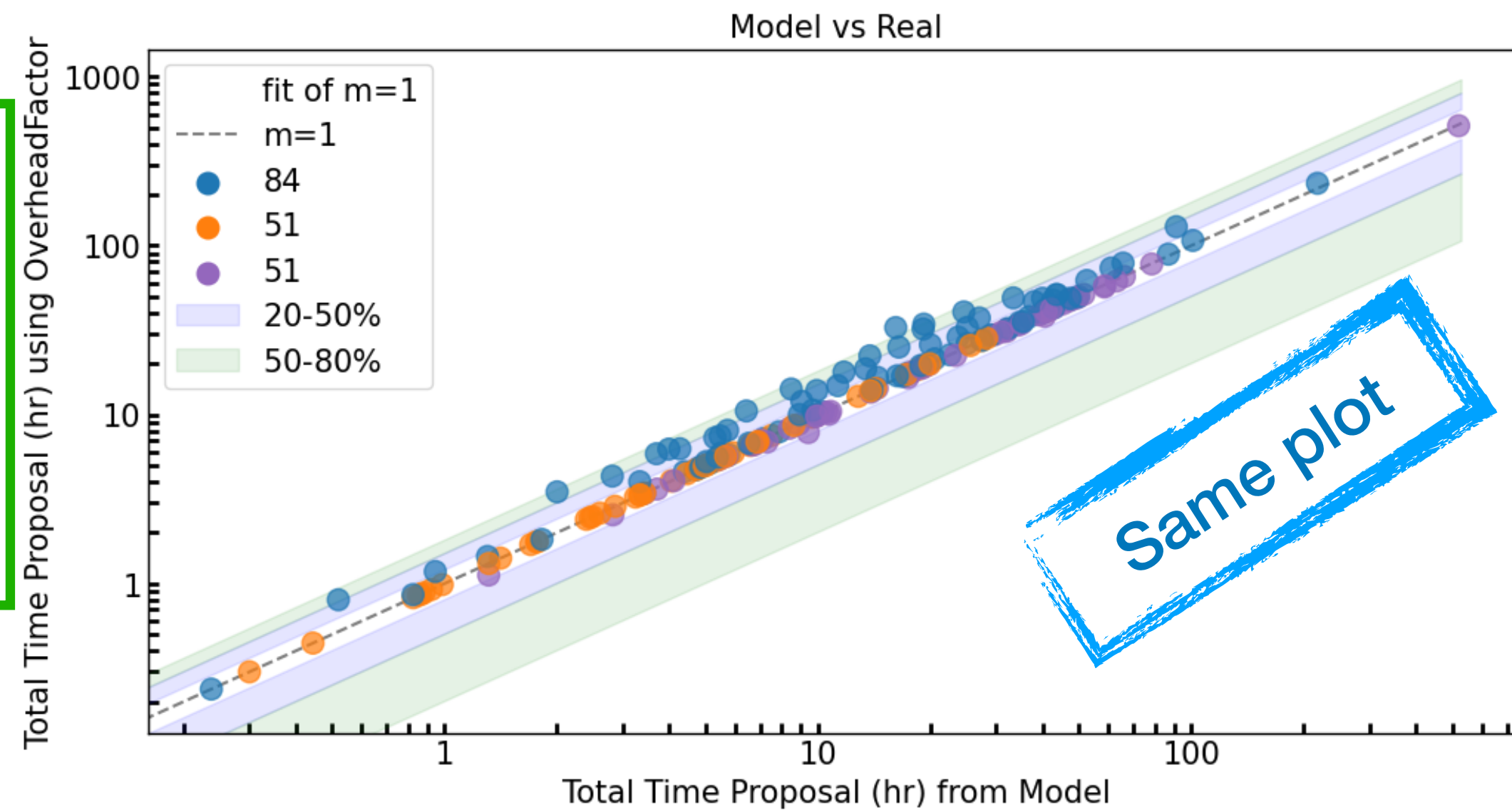
“Real” Data

- The deviation from the line is due to the number of clusters associated with the proposal.

How much Total Time does the **model** give to a proposal compared to the **real proposal**, when using the same Calibration overhead factor?

How many EB\* does the **model** predict compared to what the **real proposal**?

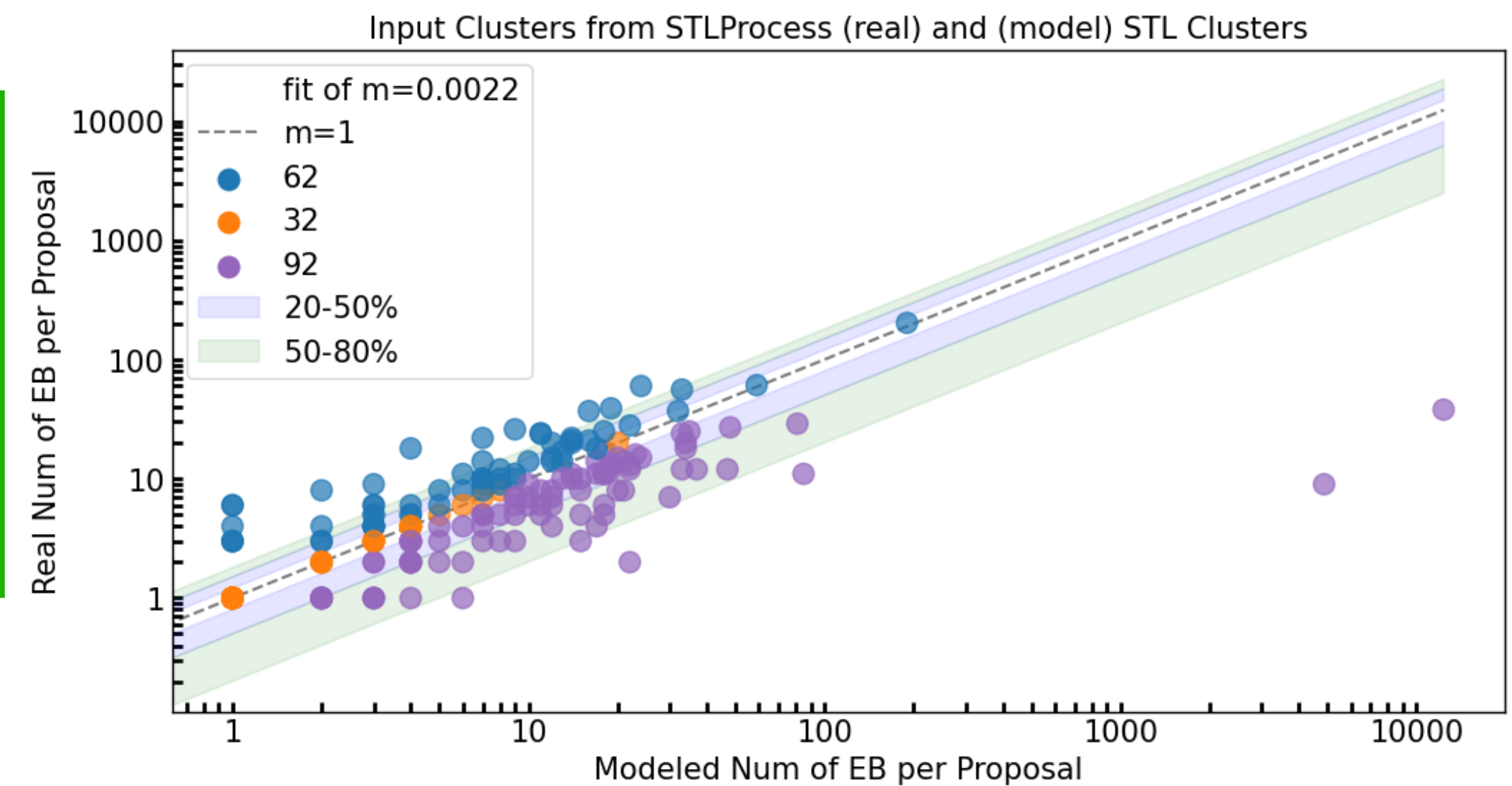
“Real” Data



“Model” Data

- The deviation from the line is due to the number of clusters associated with the proposal.

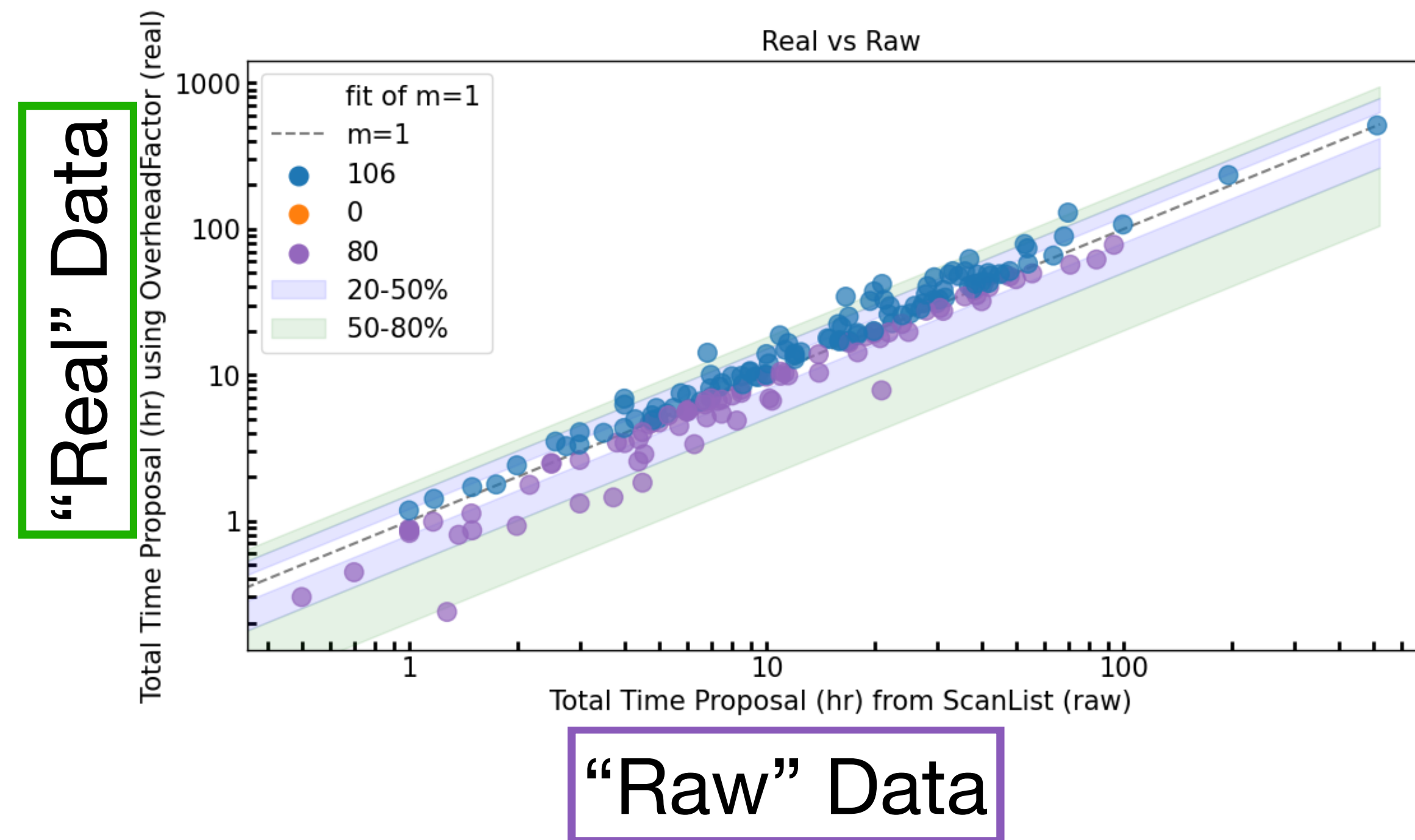
“Real” Data



“Model” Data

- The model produces more EB than the real proposal. This is likely due to restricting the total time to < 4.5 hours for a cluster in the model.
- The model tends to produce more EB but shorter EB than a real proposal

How much Total Time does **real proposal** predict with the model's **Calibration overhead factor** compared to the **raw Scan List**?

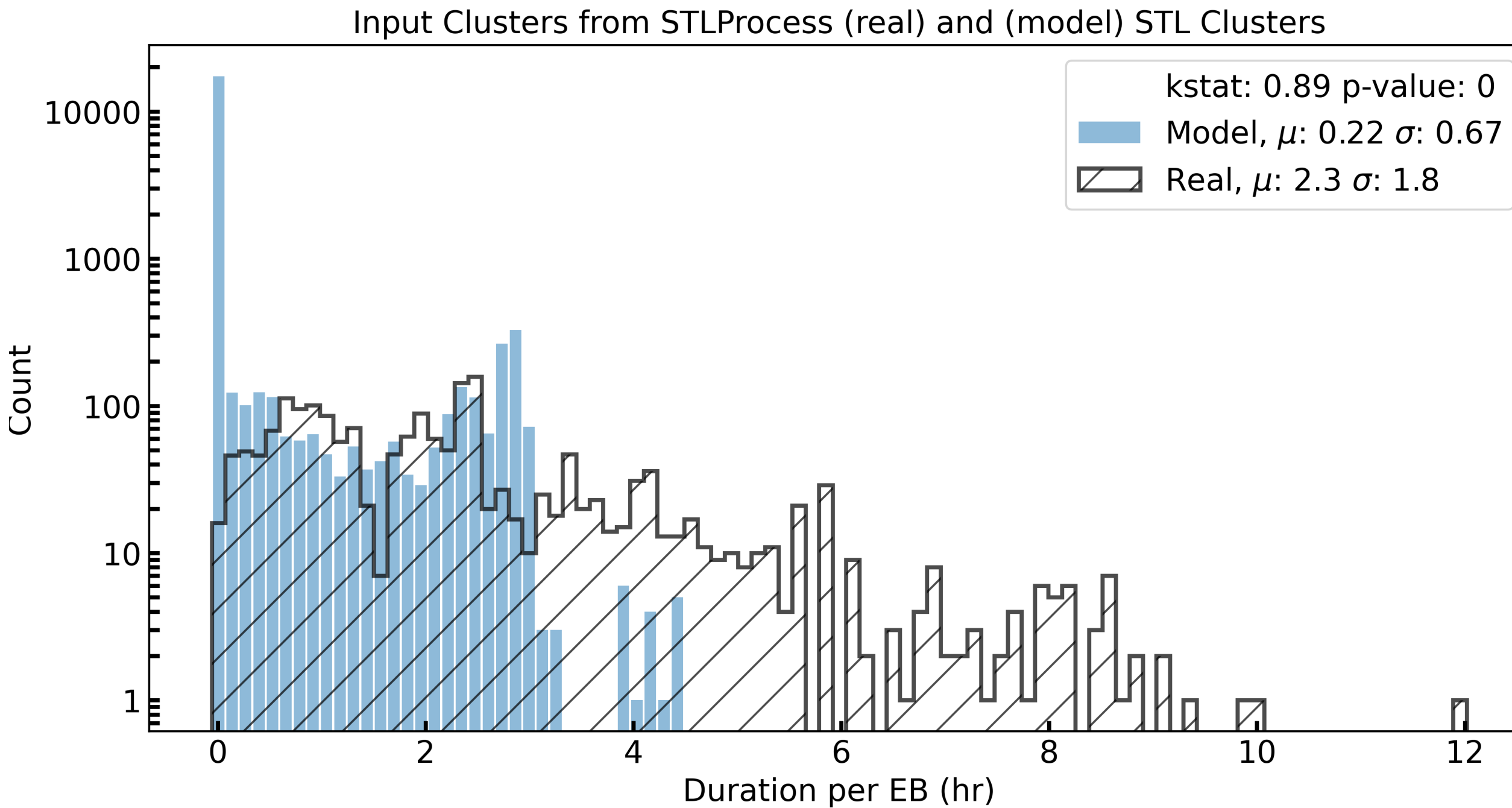


- The calibration overhead factor underestimates the total time at small requests and overestimates the total time needed at larger values. This behavior is expected.

## Duration of an EB

**“Model” Data**

~~“Real” Data~~

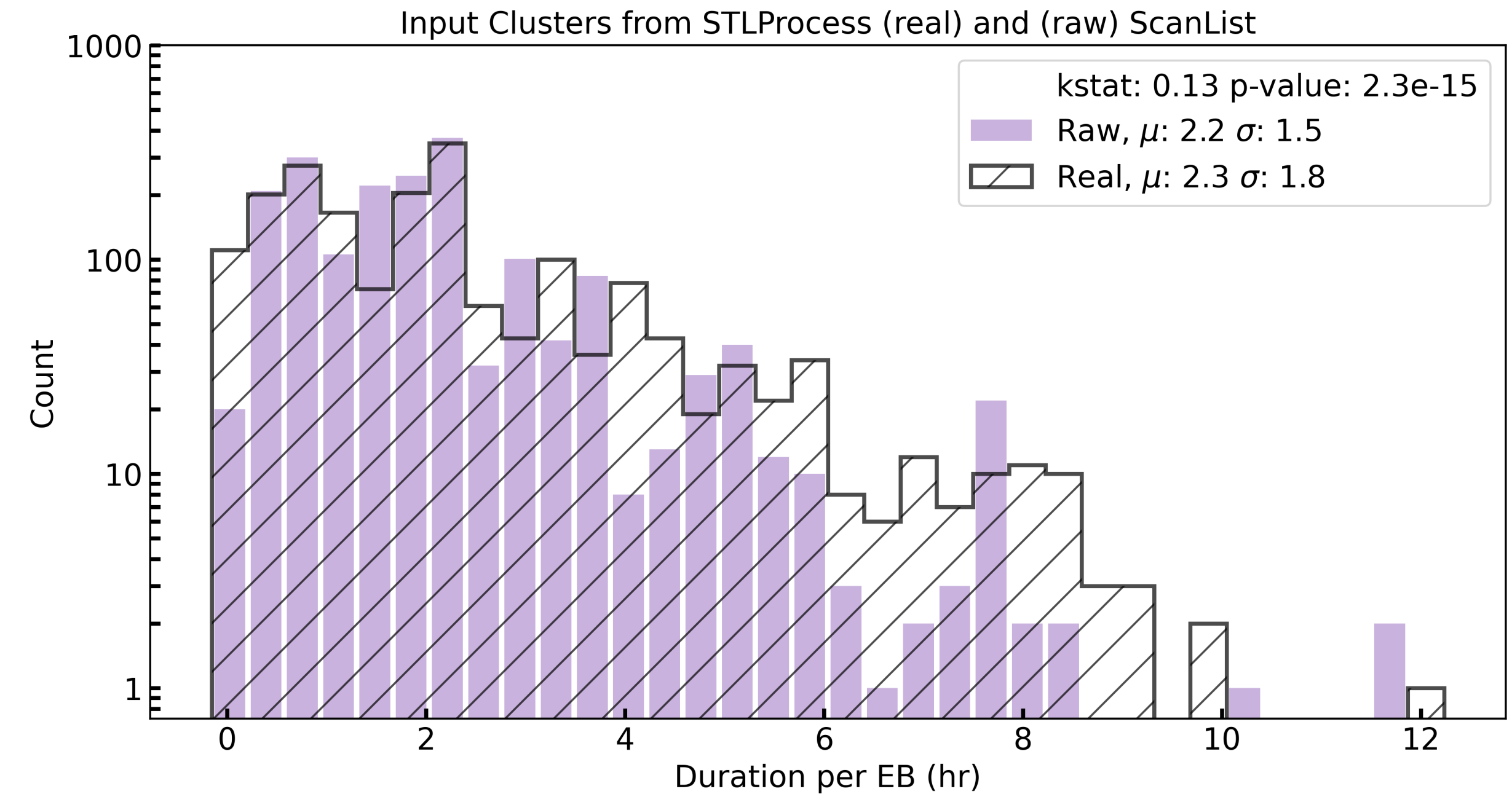


- The model is restricted from making clusters > 4.5 hours and prioritizes clusters ~ 3 hours.
- The real EB has a wide range of durations. Some of the larger EBs may have valid science motivations that the model cannot yet capture.

## Duration of an EB

~~“Real” Data~~

**“Raw” Data**



- The difference between these two populations is the Calibration Overhead Factor.

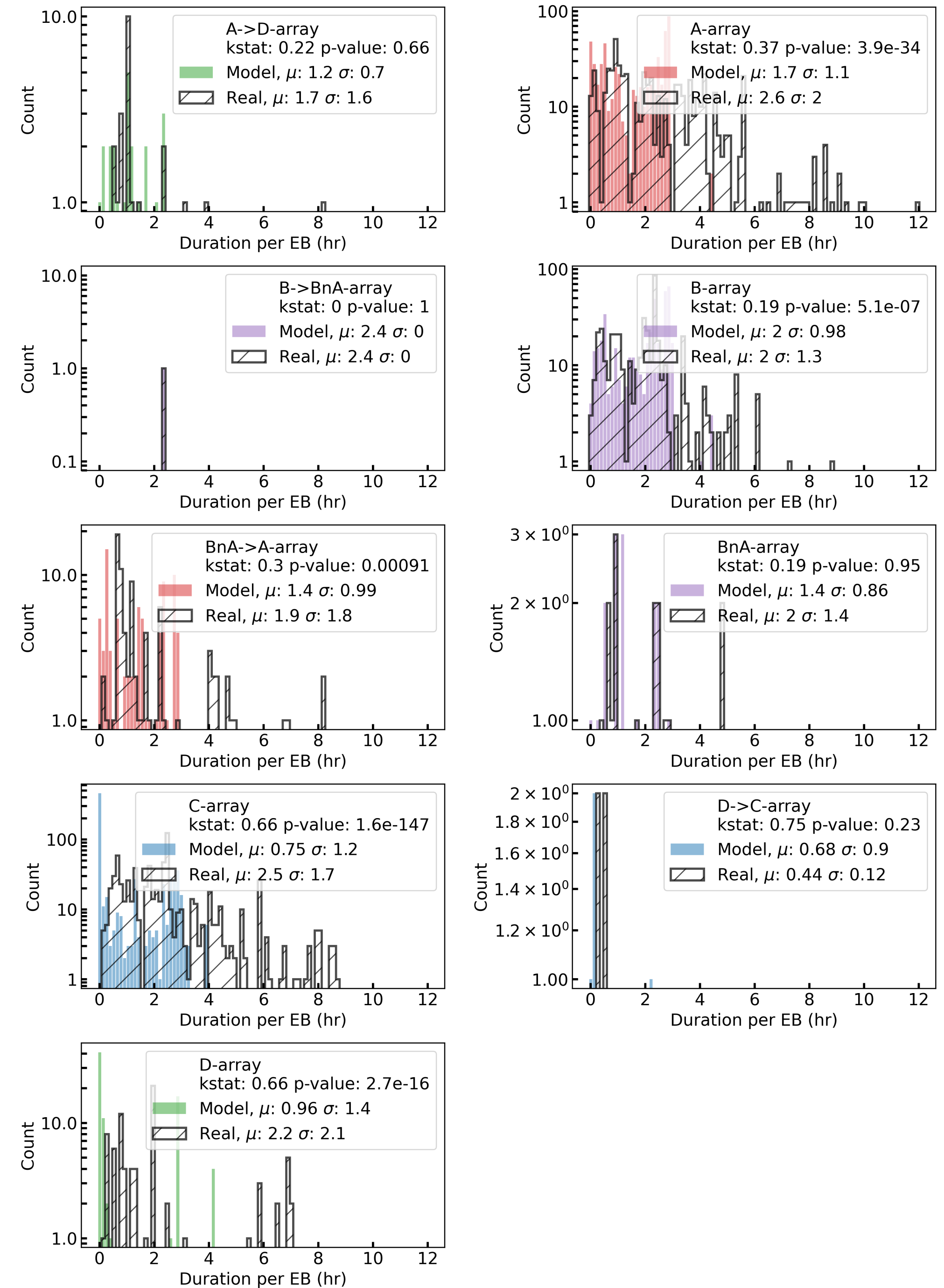
# Duration of an EB

“Model” Data

~~“Real” Data~~

- Break down of duration of EB by Configuration
- Model makes much shorter EBs in all cases than a user chose to do.

Input Clusters from STLProcess (real) and (model) STL Clusters



# Duration of an EB

~~“Real” Data~~

“Raw” Data

- Break down of duration of EB by Configuration
- The difference in population is the Calibration Overhead Factor
- The Calibration Overhead Factor might underestimate the time needed for A-array

Input Clusters from STLProcess (real) and (raw) ScanList

