Presenter Background

- Data scientist at the University of Wisconsin-Madison
    - Research focus on cosmic-ray science
    - Distributed computing

- Active in IceCube since 2004
    - Systems programmer
    - Simulation Production Coordinator
    - Data Products Coordinator
    - Core Software Team

- IceCube/HAWC PhD 2018 - Universidad de Guadalajara

Outline

- Offline Data Processing

- Simulation Production

- Public Data Releases

- Computing Requirements
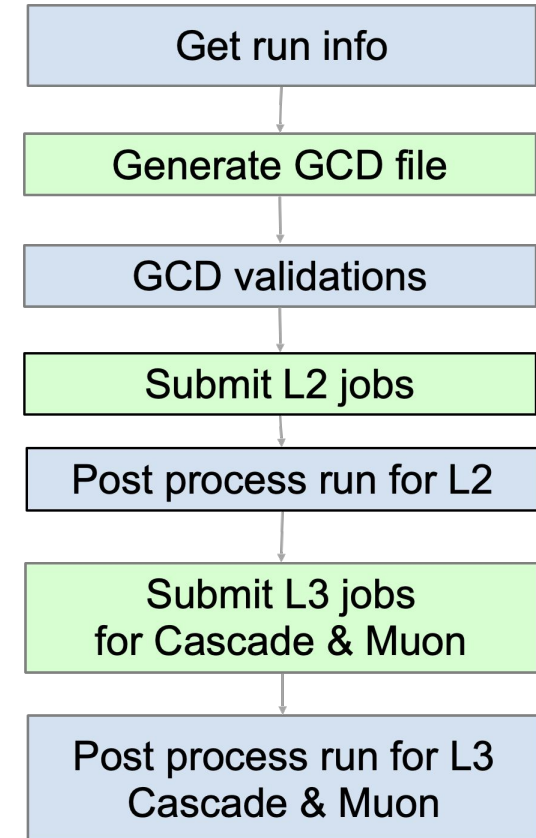
Offline Data Processing

Offline Processing (pre-2023)

Different processing levels yield increasingly specific data for different types of analyses, starting from Level1 at South Pole.

Once the data arrives via satellite,

- Level 2 processing is centrally managed (M&O) includes a number of energy and directional reconstructions.
  - Coordination with run coordinator ensures data-quality.
- Level 3 filters data specific for different analyses.
  - WG's design and validate L3 scripts
  - M&O incorporates L3 scripts in production chain to optimize resource utilization
- Data validation to detect potential issues with data value and file corruption.
- All processing is performed using IceProd and provided as a service for the Collaboration.
- Replication of all the data at the DESY–Zeuthen data center.

Workflow

Get run info

↓

Generate GCD file

↓

GCD validations

↓

Submit L2 jobs

↓

Post process run for L2

↓

Submit L3 jobs
for Cascade & Muon

↓

Post process run for L3
Cascade & Muon

Performance and Resource Utilization (pre-2023)

- Level2 clean up of obsolete filters and reconstructions reduced CPU (and memory) utilization by 36%

- Level-2 processing processing latency: elapsed time from trigger to L2 completion for 80% of the files.

- Latency for L2 reduced from ~50 days in 2018 to ~8.5 days in 2021 but increased to ~16 days in 2022-2023 due to problems with filesystem and update to the database that triggered the reprocessing of many runs.
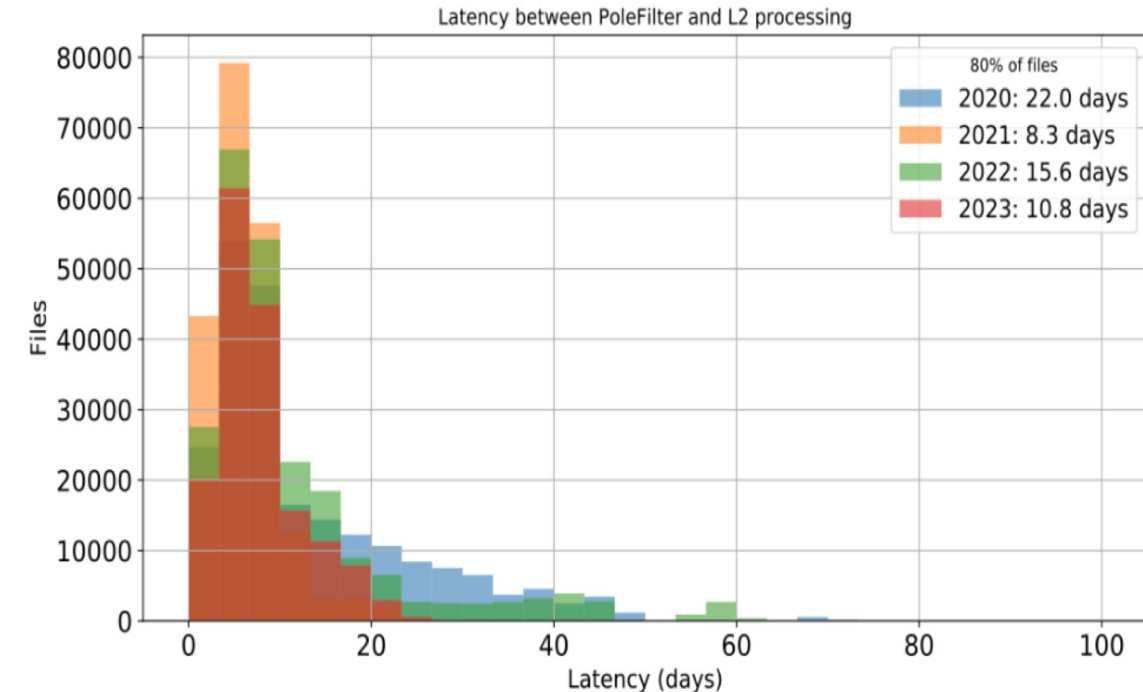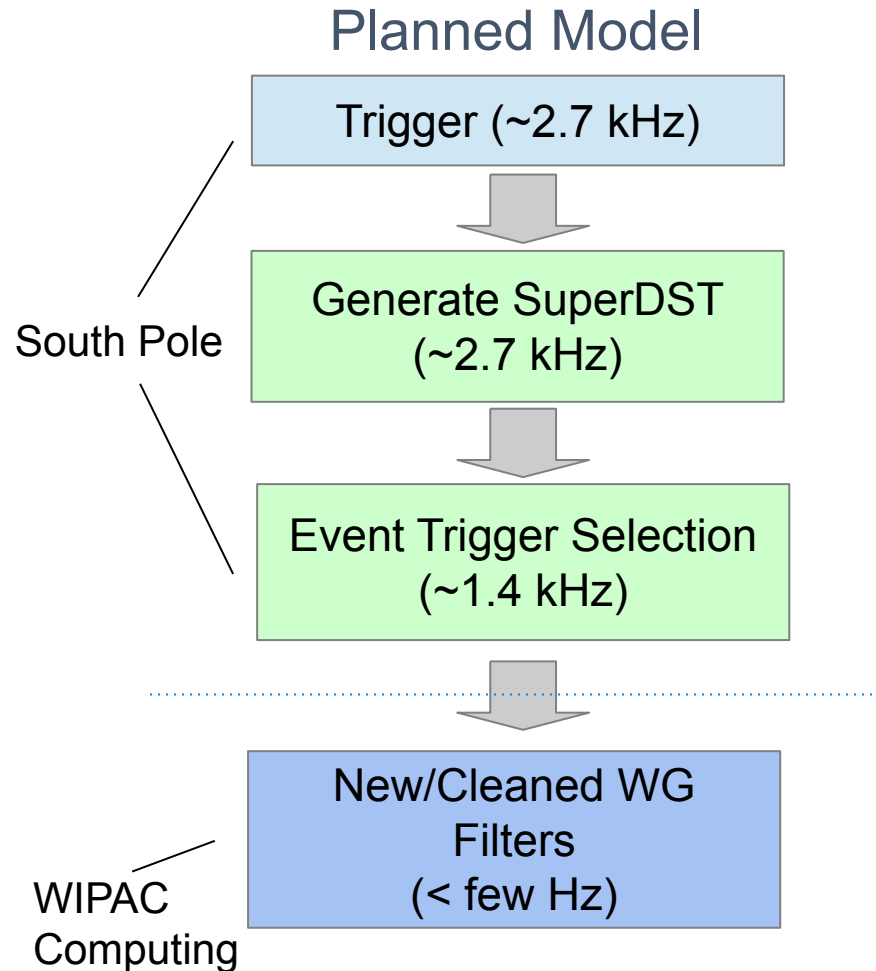

L2-processing latency distribution for the last four years. We define latency as the elapsed time between trigger to the time when L2 files are available on disk in the North. Minimum latency is determined by the weekly data validation process in the North.

L2 processing latency	< 2 weeks	10.8 days	80% quantile time from event in ice to L2 processed file in the data warehouse

New Processing: Overview

Planned Model

```
Trigger (~2.7 kHz)
        ↓
Generate SuperDST
(~2.7 kHz)
        ↓
Event Trigger Selection
(~1.4 kHz)
        ↓
New/Cleaned WG
Filters
(< few Hz)
```

South Pole

WIPAC Computing

- Pole:
  - Same DAQ triggers as always
  - Use new wavedeform (removes spurious early pulses)
  - Create SuperDST (compressed waveforms)
  - Drop Volume, String, <SMT12 triggers; Tag SMT12+ triggers with Keep_SuperDST_23 filter
  - Daily SuperDST data transfer to the North
- North:
  - Simplify and streamline filtering
  - Because new filtering was not yet ready for 2023 run: apply existing L1 & L2 in the North using modern software to validate online changes
  - New filtering to be developed & tested by WGs for use at May 2024 run start

Transitioning to new processing model

Our targeted date of April 14, 2023 for capturing requirements for the IC86–2023 run was delayed due to ongoing effort to shift to new offline filtering scheme.

Delays in the design and testing of individual filters ultimately resulted in postponement to the 2024 season.

The IC86-2023 physics run started on November 11, 2023 with run number 138615.

Stop gap: apply existing L1 & L2 in the North using modern software to validate online changes

New Milestones:

● November 2023: Started run with new online processing

● Now: New offline processing coming together

● March-April 2024: Testing and verification of new offline filters by working groups

● May 2024: New offline processing finalized
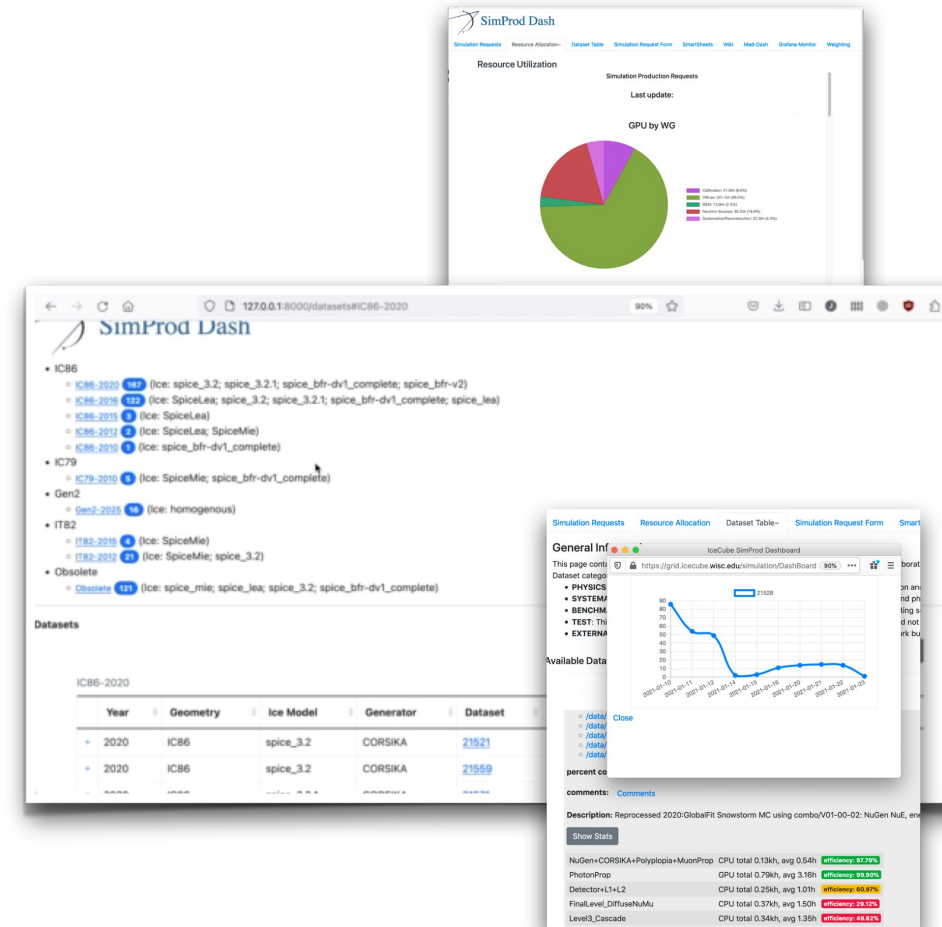
Simulation Production

Monte Carlo Simulation Production

- Simulations are required for developing analysis methods to identify signal from background, for testing the performance of reconstruction algorithms, and for determining the background contamination of data analysis samples.

- The Monte Carlo chain: background/signal events generation, propagation, detector response through L2 processing.

- Direct photon propagation on **GPU** at IceCube sites and opportunistic grid computing.

- Systematic uncertainties: Continuous variation of nuisance parameters such as, optical properties of bulk ice, DOM eff. and ang. acceptance, etc.
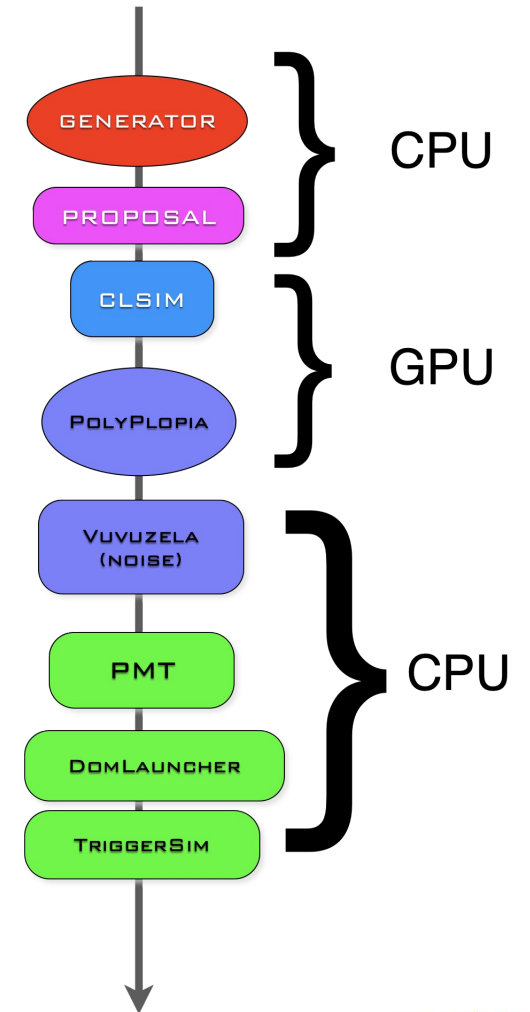
Monte Carlo Simulation Production

- Sim-Prod model transitioned from a centrally managed production to a coordinated model.

- WGs determine simulation requirements and computing needs.

- Dataset priorities discussed with technical leads and simulation is centrally generated through IceProd.

- Up-to-date summary web page with pending and complete requests for users to monitor the progress of their requests.

- Increasing number of dedicated and opportunistic resources + code optimizations.

- Development of monitoring tools to keep track of efficiency and further optimizations.

MuonGun: parametrization of atm. μ flux

- Alternative to full CORSIKA simulation

- Final-state muons weighted to parametrized μ-flux calculated from CORSIKA simulations.

- Significantly more efficient to produce.

- Need to be validated against (smaller CORSIKA datasets).

- New (re)parametrization of the atmospheric muon flux in the deep ice with post LHC models (SIBYLL2.3d)

- Expected completion by Summer 2024
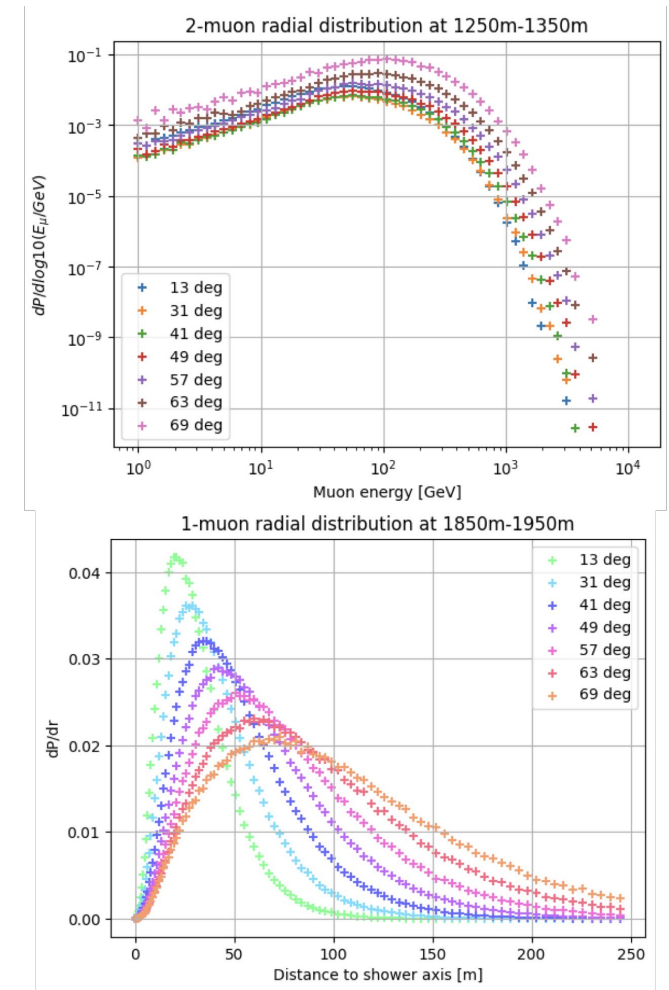
Simulation Milestones and Performance

- We expect the need to periodically re-simulate **signal** datasets with new improvements in software and modeling of optical properties of the ice.

  - Resimulation of 2020 datasets with spice 3.2.2 ice was completed in June 2023.

  - New ice models (ftpv1-3) have since been developed

  - Currently regenerating new signal datasets.

  - Obsolete simulations are scheduled to be removed from disk to accommodate for storage of new datasets.

WBS L3	Planned	Actual	Milestone
2.4.2	06/01/23	06/23/23	Re-simulate 2020 Datasets with new spice bfr-v2 ice model
2.4.2	01/01/24	TBD	Reparameterization of MuonGun with SIBYLL2.3d

Simulation Milestones and Performance

- New monitoring tools provide a means to track the efficiency of simulation jobs.

- Identification of common software-related failures and problematic computing nodes has resulted in improved efficiency.

- Better coordination with WGs has reduced the turnaround time for completion of new simulation requests.

- A couple of notable exceptions have been the result of delays due to software incompatibility
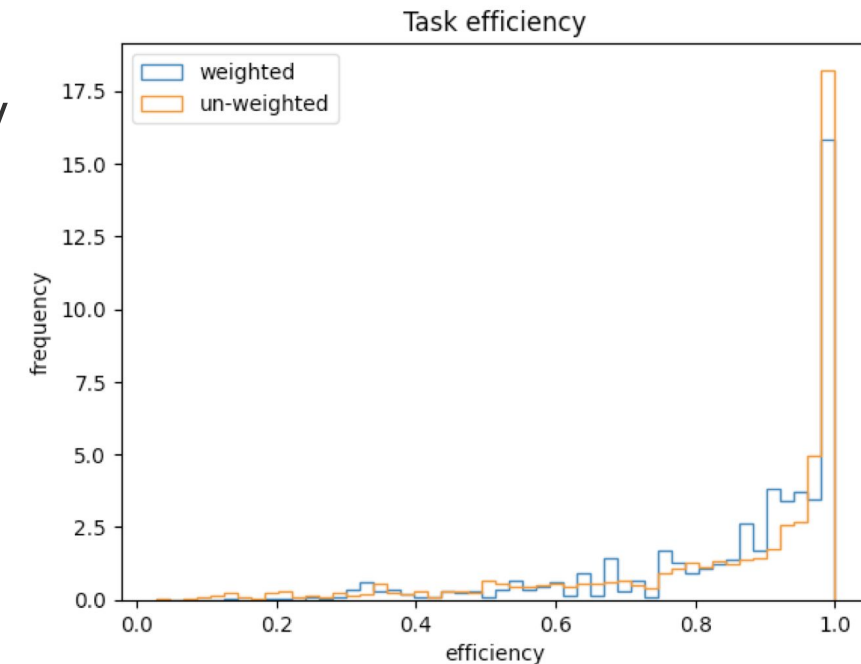
Simprod task efficiency. Blue curve corresponds to the efficiency weighted by the size of the dataset. Orange curve is the raw efficiency of tasks.

Performance Metric	Objective	Achieved	Description
Simulation Production Efficiency	> 90%	92%	Total useful time (completed jobs) divided by total computing time
Simulation Requests	< 60 days	57.2 days	90% quantile request to production

Public Data Releases

IceCube Open Data Services

Current public data consists of targeted data releases associated with specific publications such as the observation of TXS 0506+056 and the high-energy starting event selection.

- Current data releases (icecube.wisc.edu/science/data-releases) include:

  - 10 yr PS track sample - great tool for community to search for MMA correlations, and for undergrad students to work with for "starter" IceCube analyses

  - ICeCat-1 - 12+ yrs of track alert-level events from our realtime system

  - HESE 12 yr event catalog

  - Public data releases in support of most of our high profile publications

Dataverse - Harvard hosted system

(https://dataverse.harvard.edu/)

There are several sites out there aimed at making sharing of research data easier. They provide nicer ways to share, advertise, version, make findable via search, and automatically provide DOI for your data
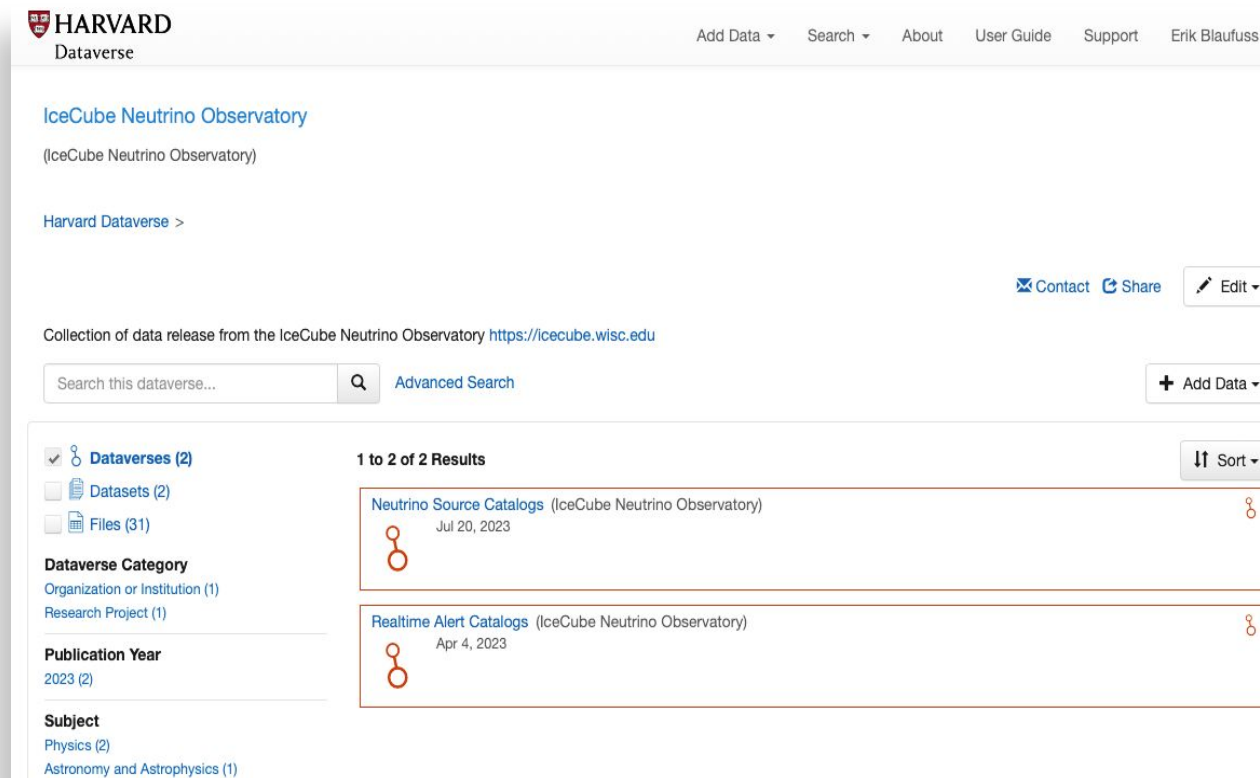
We found the Harvard Dataverse to be the most feature filled.

- Dataverse is the software that can be used to host data archives

    - Supported by Sloan, NSF, NIH, Harvard….a broad group.

- Harvard hosts one of largest free to use, publicly available instances

- Can setup data releases into organized structures (called "Dataverses"), including for named top level organizations

- the 'IceCube Neutrino Observatory' Dataverse

    - https://dataverse.harvard.edu/dataverse/icecube

IceCube Dataverse

Recent data releases in nu-sources (HESE 12 year) and realtime (IceCat-1 alerts)

All future data releases, and some of the widely used historical data sets from our website (e.g. 10 yr PS release, etc) to be imported into Dataverse

Computing Requirements

Computing Needs

- Offline production utilizes ~480k CPU-hours/year.

- Filter cleanup resulted in 36% reduction of CPU utilization and a comparable reduction in memory requirements.

- 100 TB of storage per year for both the Pole-filtered input data and output data from offline production.

- **Pass2** utilized 10.9M CPU-hours and 520 TB storage.

- Simulation requirements are dominated by background CORSIKA.

- Background simulations for the in-ice array require roughly 30k years of CPU time and about 2.4k years of GPU time to produce and filter.

- Anticipate a future **Pass3** will require ~20M CPU-hours.

- In addition to IceTop surface array simulations and signal simulation.

- Anticipated increase in needs due to complexity from IceCube Upgrade.

CORSIKA Benchmarks
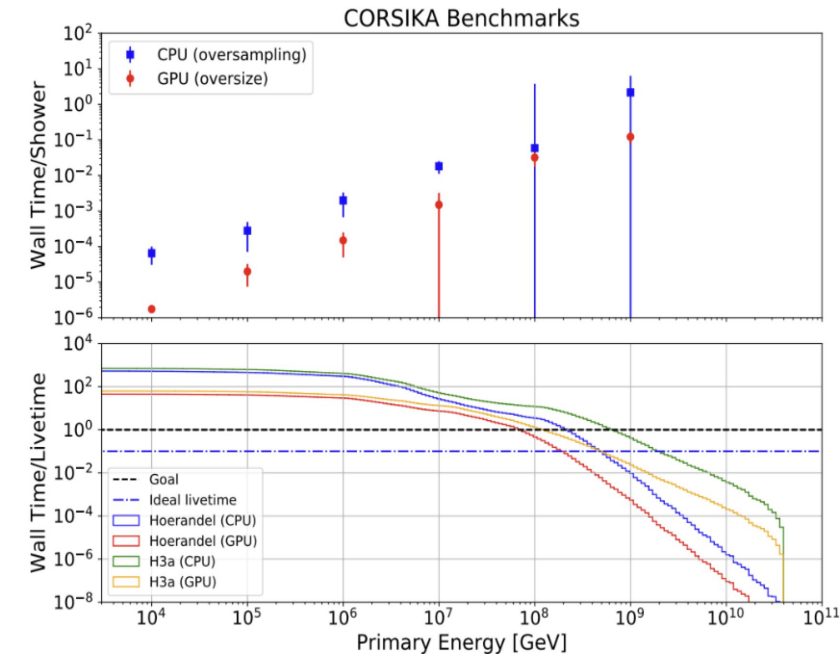
Summary & Conclusions

- With nearly all events transferred in SuperDST format, we are able to move a lot of the filtering to the North.

  - This is an opportunity to **simplify and streamline filtering.**

  - New filtering to be developed & tested by WGs for use in 2024 run start.

- Background MC is always challenging but parametric tools and importance sampling helps to bridge the gap for analyzers.

  - New SimProd webpage provides a detailed catalog of MC datasets

  - MC datasets are generated per request and discussed with WG to better serve the Collaboration.

  - Improvements in software and increasingly more accurate ice models require periodic need for re-simulating MC.

- We recently adopted Dataverse to host public data archives.

  - Older data releases will remain available on current site.

Backup

Pass2+ reprocessing (2010-2016)

- Improvements in software and detector calibration required that we reprocess nearly 10 years of data.

- This involves retrieving data from long-term archive and new workflow combining online and offline processing.

- Pass2 provided unified data set and included a total of seven years of data (2010–2016).

- The reprocessing campaign lasted 13 months.

- Anticipate additional data reprocessing (Pass3+) campaigns as further improvements in calibration and feature extraction are developed.

- Improvements in calibration resulting from the IceCube Upgrade will certainly require additional re-processing campaigns. We anticipate at least two additional reprocessing campaigns.

Current Background Simulation

- Individual datasets for targeted analyses.

- Datasets configured with different energy spectra and energy ranges.

- Obsolete simulations removed.

- For lower energies we approximate natural spectrum with E^-2.65

- For energies above 10^6 GeV, we simulate a harder spectrum with E^-2
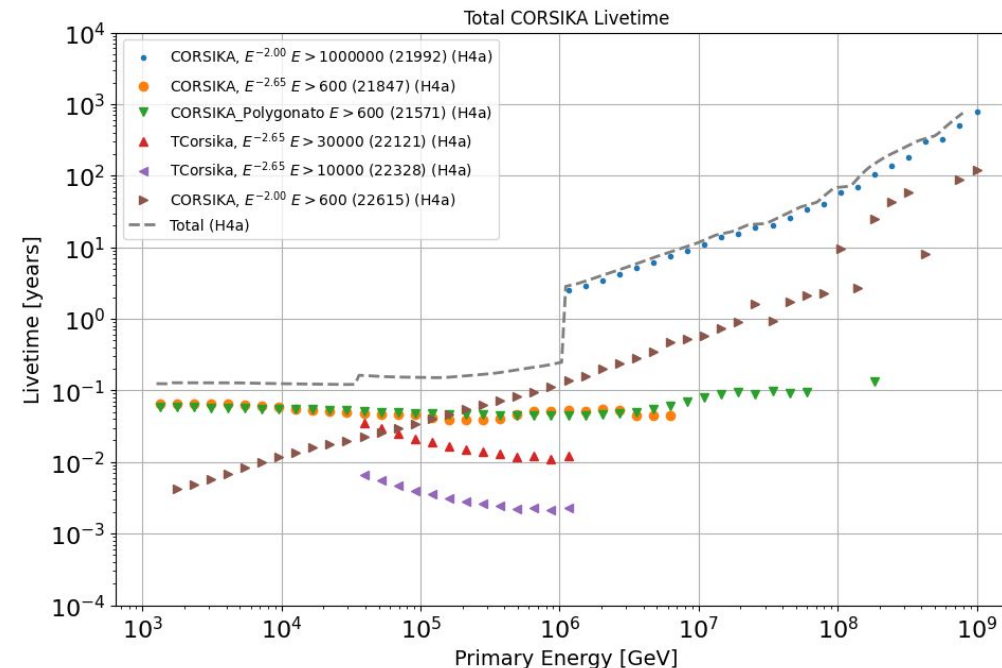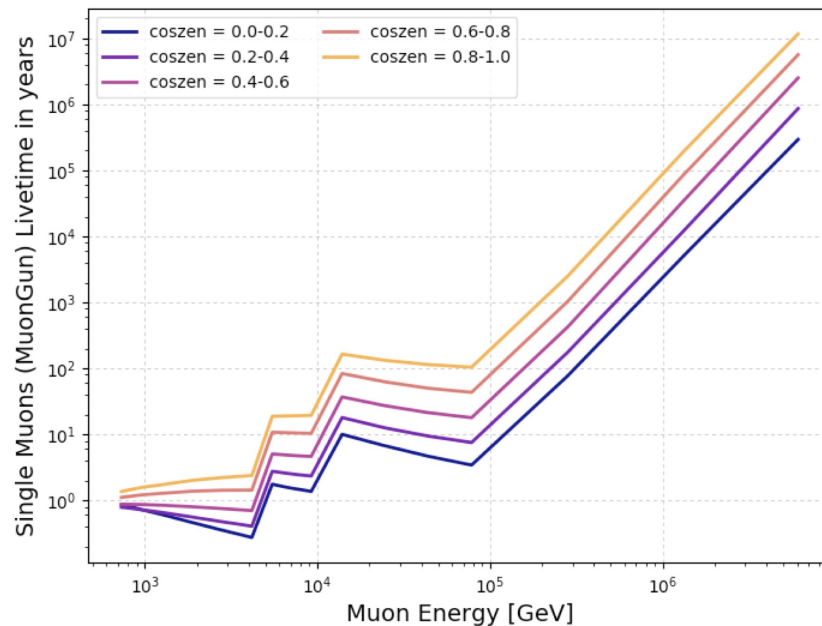
Current Signal Simulation

- Individual datasets for targeted analyses.

- Datasets configured with different energy spectra and energy ranges.
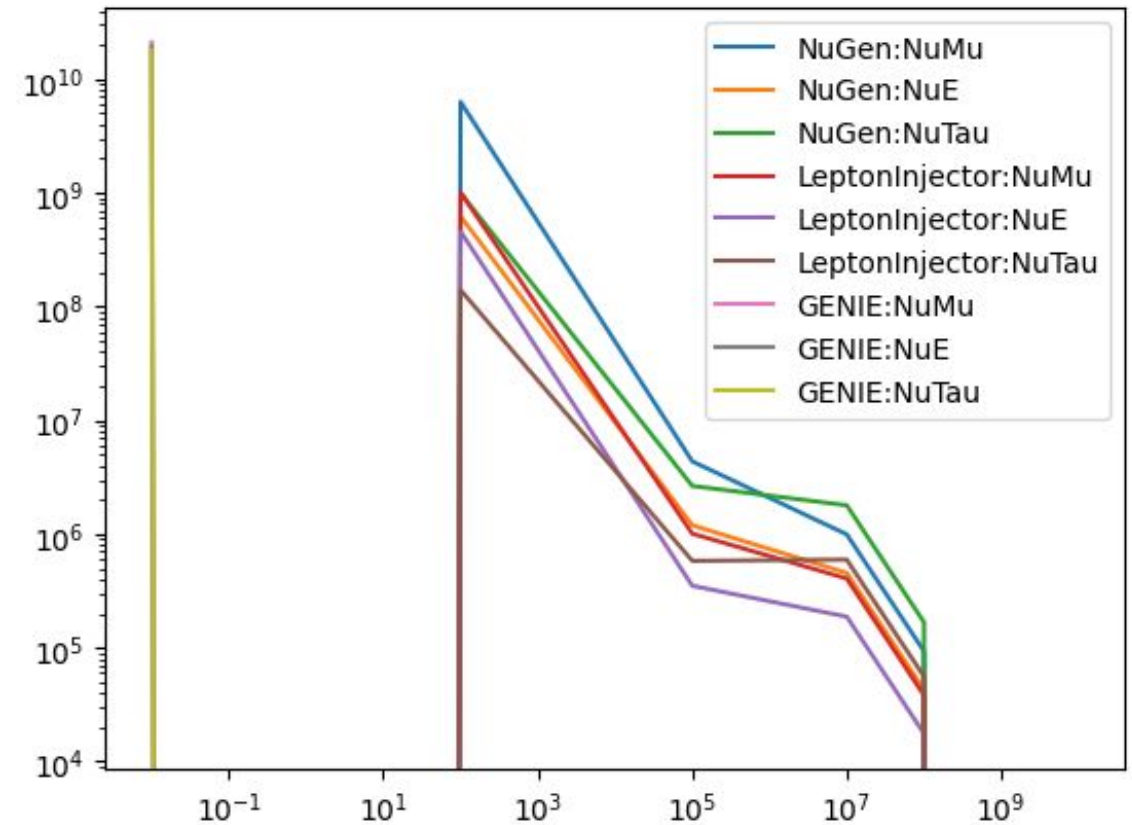
- Obsolete simulations removed.

Simulation Products for PY3

Energy (GeV)	1.0×10^{-2}	1.0×10^{2}	1.0×10^{5}	1.0×10^{7}	1.0×10^{8}	1.0×10^{9}	1.0×10^{10}
NuGen (events)							
NuMu	0.0	8.6×10^{8}	7.6×10^{5}	1.9×10^{5}	1.8×10^{4}	0.0	0.0
NuE	0.0	6.4×10^{7}	2.3×10^{5}	8.3×10^{4}	7.8×10^{3}	0.0	0.0
NuTau	0.0	1.2×10^{8}	4.6×10^{5}	3.4×10^{5}	3.2×10^{4}	0.0	0.0
LeptonInjector (events)							
NuMu	0.0	1.0×10^{9}	1.0×10^{6}	4.1×10^{5}	3.8×10^{4}	0.0	0.0
NuE	0.0	4.6×10^{8}	3.5×10^{5}	1.9×10^{5}	1.8×10^{4}	0.0	0.0
NuTau	0.0	1.4×10^{8}	5.8×10^{5}	6.0×10^{5}	5.6×10^{4}	0.0	0.0
GENIE (events)							
NuMu	3.9×10^{8}	0.0	0.0	0.0	0.0	0.0	0.0
NuE	3.0×10^{8}	0.0	0.0	0.0	0.0	0.0	0.0
NuTau	5.2×10^{8}	0.0	0.0	0.0	0.0	0.0	0.0
Background livetime (days)							
CORSIKA	0.0	8.0×10^{0}	8.7×10^{0}	6.8×10^{1}	7.4×10^{2}	0.0	0.0
MuonGun	0.0	0.0	1×10^{-1}	1×10^{-1}	0.0	0.0	0.0

TABLE 1. PY3 Resource Utilization

Generator	cpu (kh)	gpu (kh)	storage (TB)
CORSIKA-in-ice	1.8×10^{3}	3.8×10^{2}	9.9×10^{1}
NuGen	6.3×10^{2}	3.8×10^{2}	4.2×10^{1}
CORSIKA-icetop	1.7×10^{3}	6.6×10^{0}	–
GENIE	4.1×10^{2}	8.5×10^{0}	3.6×10^{-2}
LeptonInjector	2.2×10^{2}	1.7×10^{2}	–
MuonGun	1.6×10^{2}	1.5×10^{2}	–

Resource utilization: storage information not available for Lustre file system beyond 2020

Simulation generated during PY3 as function of energy, including neutrinos (NuGen, LeptonInjector, and GENIE) and cosmic-ray background (CORSIKA, MuonGun). Background generation was suspended for the greater part of 2023 due to storage limitations.

Current MC Catalog

Energy (GeV)	1.0×10^{-2}	1.0×10^{2}	1.0×10^{5}	1.0×10^{7}	1.0×10^{8}	1.0×10^{9}	$1.0 \times 10^{+10}$
NuGen (events)							
NuMu	0.0	6.3×10^{9}	4.4×10^{6}	9.9×10^{5}	9.3×10^{4}	0.0	0.0
NuE	0.0	6.2×10^{8}	1.2×10^{6}	4.5×10^{5}	4.3×10^{4}	0.0	0.0
NuTau	0.0	1.0×10^{9}	2.6×10^{6}	1.8×10^{6}	1.7×10^{5}	0.0	0.0
LeptonInjector (events)							
NuMu	0.0	1.0×10^{9}	1.0×10^{6}	4.1×10^{5}	3.8×10^{4}	0.0	0.0
NuE	0.0	4.6×10^{8}	3.5×10^{5}	1.9×10^{5}	1.8×10^{4}	0.0	0.0
NuTau	0.0	1.4×10^{8}	5.8×10^{5}	6.0×10^{5}	5.6×10^{4}	0.0	0.0
GENIE (events)							
NuMu	$2.1 \times 10^{+10}$	0.0	0.0	0.0	0.0	0.0	0.0
NuE	$1.9 \times 10^{+10}$	0.0	0.0	0.0	0.0	0.0	0.0
NuTau	$1.8 \times 10^{+10}$	0.0	0.0	0.0	0.0	0.0	0.0
Background livetime (days))							
CORSIKA	0.0	9.8×10^{2}	9.8×10^{2}	4.0×10^{3}	2.3×10^{4}	1.5×10^{5}	0.0
MuonGun	0.0	0.0	2×10^{0}	1×10^{1}	1×10^{5}	0.0	0.0

Simulation in current MC catalog as function of energy, including neutrinos (NuGen, LeptonInjector, and GENIE) and cosmic-ray background (CORSIKA, MuonGun).

TABLE 2. Total for Current Catalog

Generator	cpu (kh)	gpu (kh)	storage (TB)
CORSIKA-in-ice	2.4×10^{4}	2.7×10^{3}	1.4×10^{3}
NuGen	2.5×10^{3}	3.0×10^{3}	3.5×10^{2}
CORSIKA-icetop	3.4×10^{3}	3.3×10^{1}	1.4×10^{2}
Other	6.7×10^{2}	–	1.5×10^{1}
MuonGun	7.9×10^{2}	1.1×10^{3}	5.7×10^{1}
GENIE	3.0×10^{3}	3.9×10^{2}	3.7×10^{2}
LeptonInjector	2.2×10^{2}	1.7×10^{2}	–

Resource utilization: storage information not available for Lustre file system beyond 2020