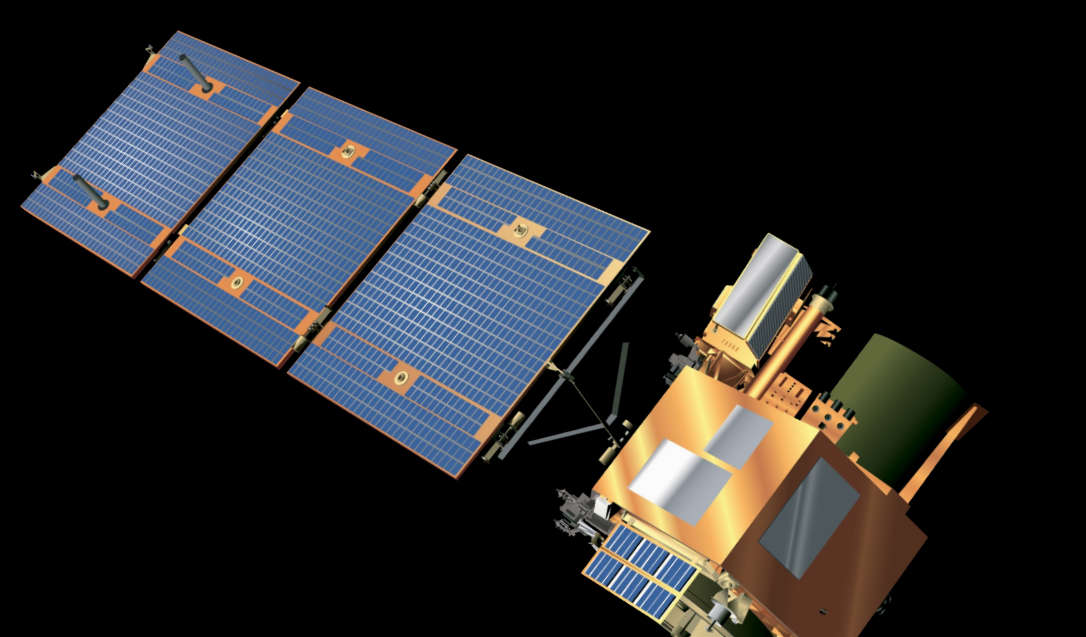
GC42D-0750: UPDATES IN DEVELOPING A PROTOTYPE SCIENCE PIPELINE AND FULL-VOLUME, GLOBAL HYPERSPECTRAL SYNTHETIC DATA SETS FOR NASA'S EARTH SYSTEM OBSERVATORY'S UPCOMING SURFACE, BIOLOGY AND GEOLOGY MISSION

Jon M. Jenkins[1], Peter Tenenbaum[1,2], Yohei Shinozuka[1,3], Bill Wohler[1,2], Weile Wang[1], Andrew Michaelis[1], Jennifer L. Dungan[1], Ian G. Brosnan[1], Vanessa Genovese[1,3], Michelle M. Gierach[5], Philip Townsend[6,5], Benjamin Poulter[7], and Adam Chlus[5]

[1]NASA Ames Research Center, [2]SETI Institute, [3]Bay Area Environmental Research Institute, [4]California State University – Monterey Bay, [5]Jet Propulsion Laboratory, [6]University of Wisconsin, Madison, [7]NASA Goddard Space Flight Center



Abstract:

The Surface Biology and Geology (SBG) mission recently passed mission confirmation review and has entered phase A – design and development. SBG will acquire high resolution solar-reflected spectroscopy and thermal infrared observations at a data rate of ~2.5 TB/day and generate products at ~40 TB/day. Given that the per-day volume is greater than NASA's total extant airborne hyperspectral data collection, collecting, processing, disseminating, and exploiting the SBG data present new challenges. To meet these challenges, we have developed a prototype science pipeline and a full-volume global hyperspectral synthetic data set to help prepare for SBG's flight (see poster GC42D-0730). Our science pipeline is based on the science processing technology developed for NASA's Kepler and TESS planet-hunting missions. The pipeline infrastructure, Ziggy, provides a scalable architecture for robust, repeatable, and replicable science and application products that can be run on a range of systems from a laptop to the cloud or a supercomputer. Ziggy is compliant with NASA Procedural Requirement (NPR) 7150.2C, is at a technical readiness level (TRL) of 7 and has been released to github.com/nasa/ziggy. We integrated Ziggy with EO-1/Hyperion workflows to build a prototype pipeline and ingested the 17-year mission archive that provides globally sampled visible through shortwave infrared spectra that are representative of SBG data types and volumes. We fully implemented the first stage and processed the entire 55 TB Hyperion data set from the raw data (Level 0) to top-of-the-atmosphere radiance (Level 1R). We are currently evaluating the ISOFIT atmospheric correction module to convert the L1R data to surface reflectance (Level 2) before reprocessing the full data set to L2. Crosschecks are being performed with RadCalNet as well as with coincident observations by AVIRIS. We are also investigating modern methods for georectifying the Hyperion scenes. Finally, we describe an analysis of the cost to conduct forward processing and reprocessing campaigns for SBG on HECC with dedicated compute and storage resources using the resurrected Hyperion pipeline as a proxy for full-volume SBG data. The analysis demonstrates that SBG L0 data can be processed to L2 on HECC with full reprocessing campaigns every two years for ~$2.6M over a 7-year lifespan. Moreover, 69% of the system capacity would be available for other activities, possibly enabling future open-source science activities, including algorithm development, L3+ processing, .etc.

1. Reanimating the Hyperion Pipeline:

We've reconstructed the Hyperion/EO-1 pipeline in Python as shown in Figure 1 and have processed the entire 55 TB data set [1] to top-of-atmosphere radiance (L1A) at full resolution (27 TB). Our pipeline allows the atmospheric correction with either the Imaging Spectrometer Optimal FITting (ISOFIT) open source package version 2.9.2 [2] or the non-open source Atmospheric Removal (ATREM) program version 5.0 (for testing and comparison purposes) [3]. We are now working to incorporate georeferencing and co-registration into the pipeline to complete the pipeline to an L1B georeferenced top-of-atmosphere radiance (approx. 27 TB), and georeferenced, orthocorrected level 2 (L2) reflectance product (approx. 124 TB). We also have begun conducting experiments to evaluate and verify the results and are investigating how the processing scales with data volume for both L0→L1A, L1A→L1B, and L1B→L2.

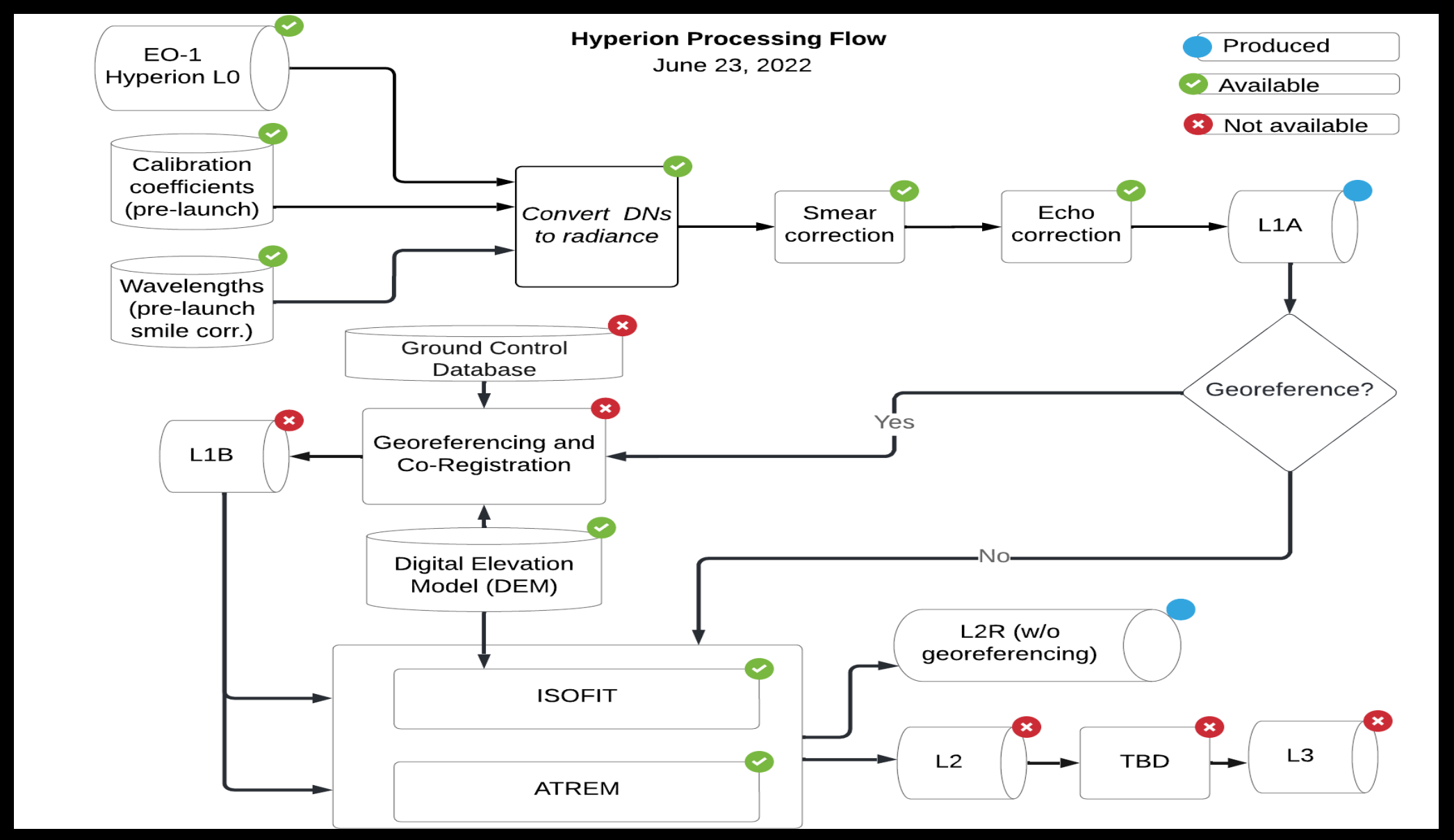


Figure 1. Hyperion Processing Flow.

2. Evaluating the Hyperion Pipeline:

We have been evaluating and verifying the results for specific scenes:

- Comparing the Hyperion L2 surface reflectance reconstructed with ISOFIT to ground-based RadCalNet observations [4b].
- Comparing the ISOFIT atmospheric products to the ground-based AERONET observations [5b].
- Comparing surface reflectance spectra generated using ATREM with those using ISOFIT.
- Comparing surface reflectance spectra of Hyperion with airborne AVIRIS scenes.
- Comparing Hyperion results with forward modeling using MODIS data.

The Hyperion reflectance spectra show some unphysical patterns, which may indicate calibration corrections not available to this workflow [6b] and are expected (Petya Campbell, personal communication). Nonetheless, they show generally good agreement with near-coincident RadCalNet observations at the Railroad Valley site (Figure 2). The root-mean-square difference across the wavelength is 0.04-0.09 for the four pairs shown in Figure 2. Atmospheric changes between each pair of measurements may explain the reflectance differences such as those noticeable in the bottom panel.

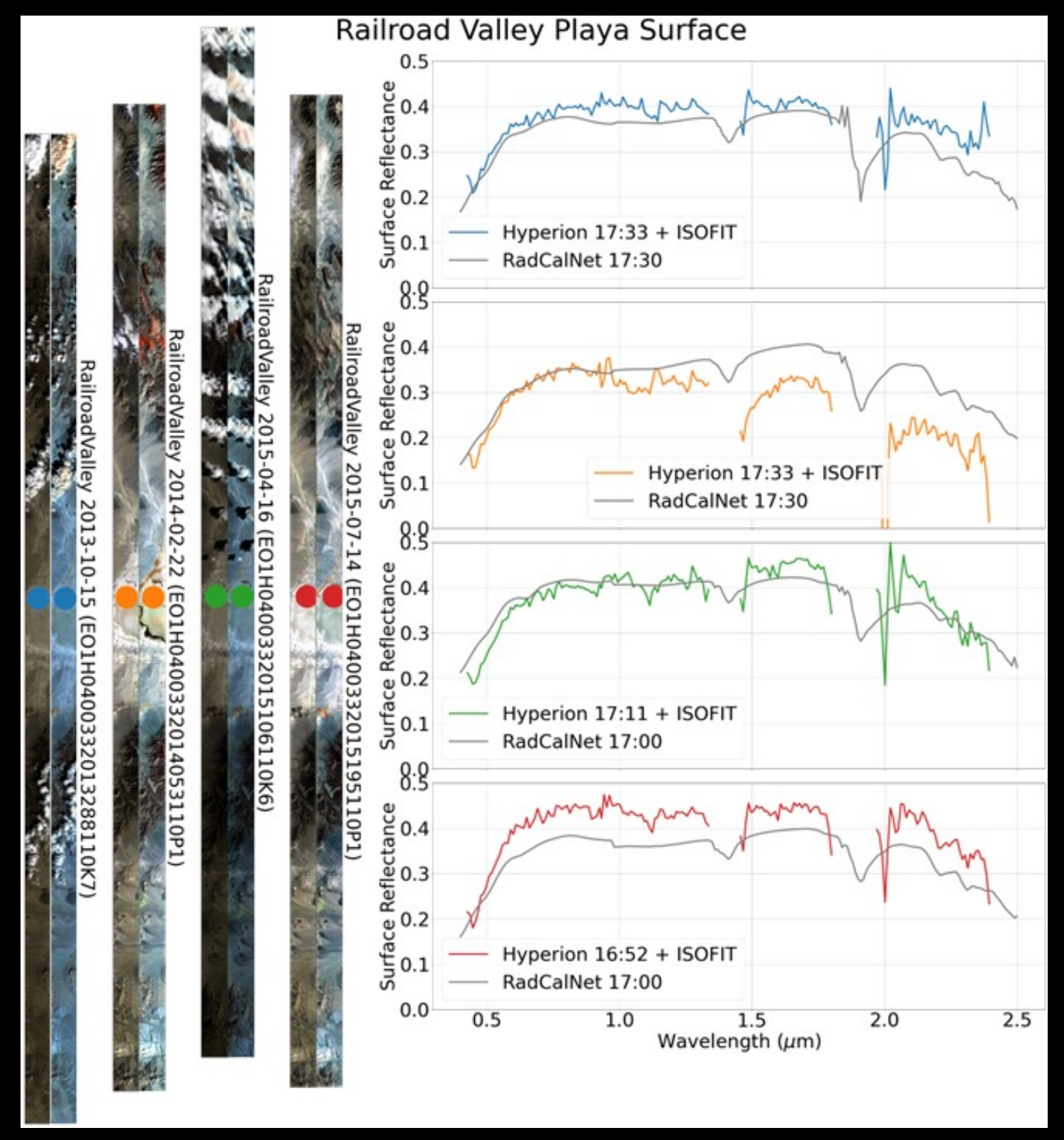


Figure 2. Comparison of RadCalNet measurements with Hyperion surface reflectance retrievals for scenes observed in Railroad Valley. Four VNIR/SWIR Hyperion image pairs appear to the left with markers indicating the pixel location for the surface reflectance plotted on the right-hand side vs. wavelength. The solid black curves are the results obtained by RadCalNet. The blue, orange, green and red curves show Hyperion results obtained using the ISOFIT package for atmospheric corrections.

3.Atmospheric Retrievals:

ISOFIT retrieves atmospheric products in addition to surface reflectance, enabling comparison of the Hyperion scenes that include Railroad Valley with the AERONET observations. The RMS differences for near-coincident cases (< 1 km horizontal distance and <30-minute temporal gap, including scenes not shown in Figure 2) marked with closed circles in Figure 3 is 0.14 cm in column water vapor and 0.11 in 500-550 nm aerosol optical depth, respectively. The aerosol optical depth retrieved with Hyperion and ISOFIT tends to be greater than the ground-observed value. Systematic differences may be explained by calibration inaccuracy; their dependence on wavelength, location and year remain to be investigated.

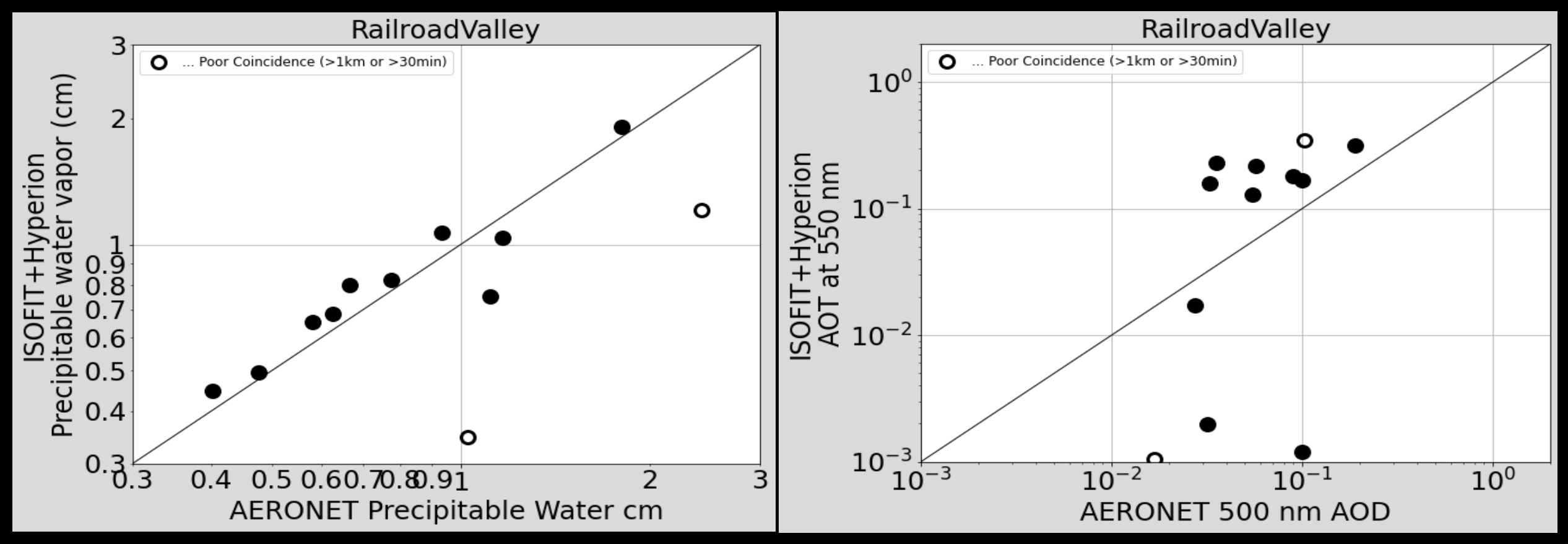


Figure 3. Comparison of Hyperion ISOFIT atmospheric products (left: water vapor, right: aerosol optical depth) with AERONET observations at Railroad Valley.

4. Comparison of surface reflectance spectra retrieved from Hyperion and AVIRIS

Figure 4 shows the retrieval products for near-coincident airborne AVIRIS-Classic observations generated with the same two algorithms. Note that georeferencing errors are expected, perhaps by 5-10 pixels. The comparison between the two instruments highlights the unphysical spectral patterns of Hyperion, such as the noisiness across the entire wavelength range and the weak signal below 450 nm. While good agreement between the two instruments is evident near 600 nm and near 1600 nm, the Hyperion reflectance is smaller than the AVIRIS in other spectral regions.

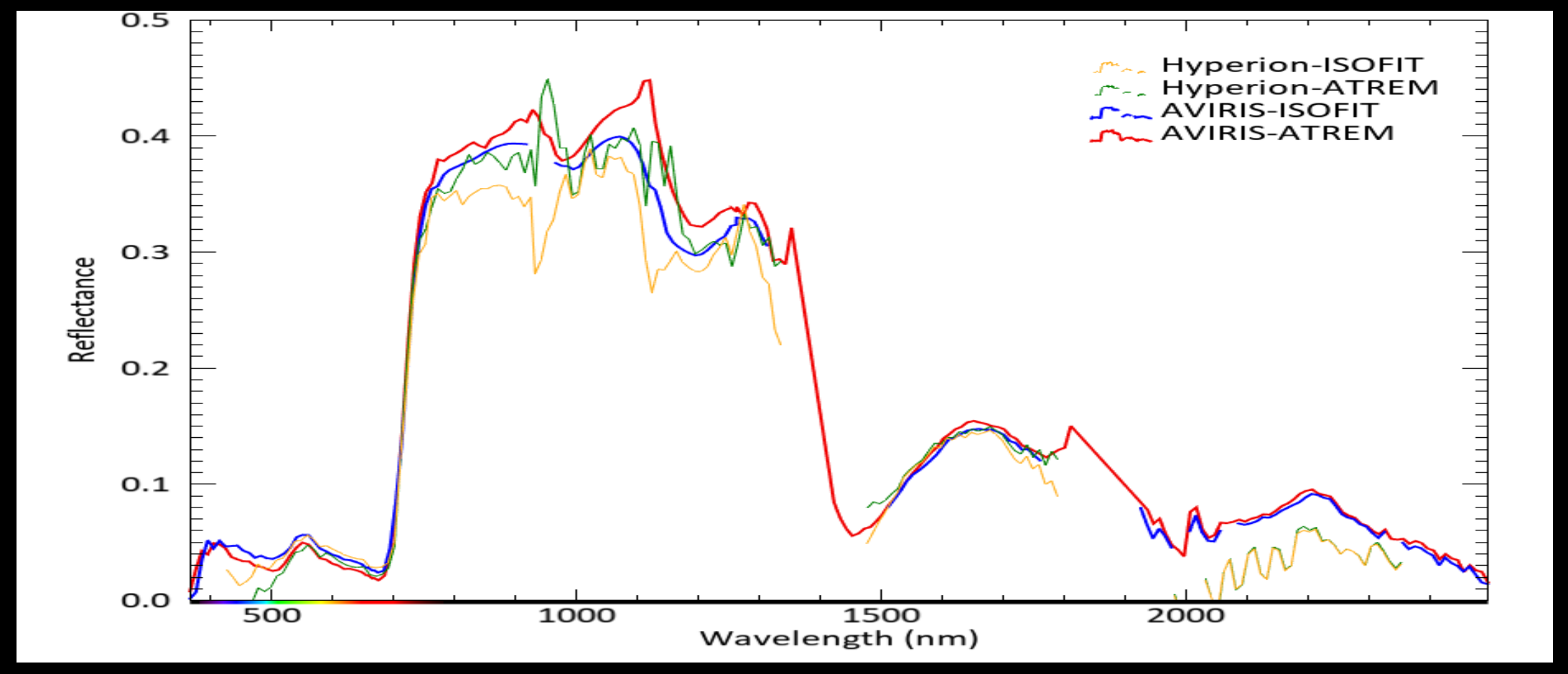
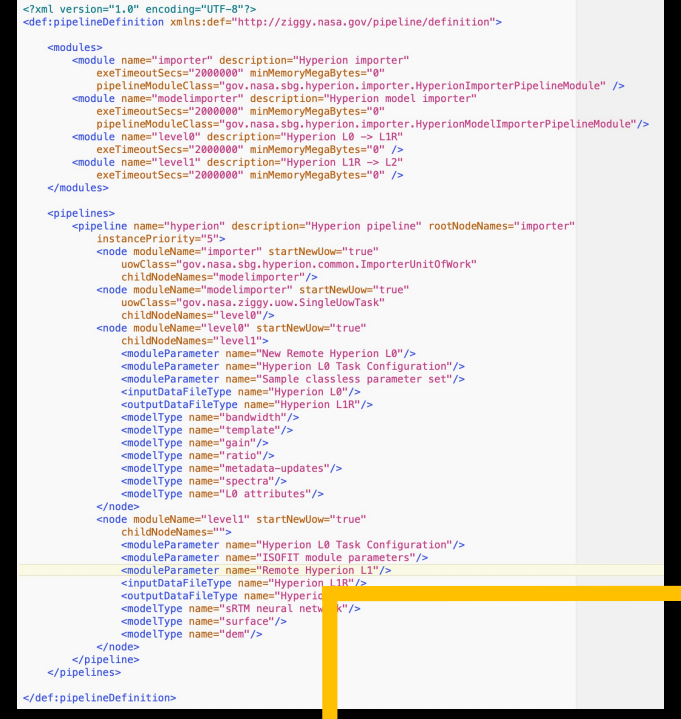


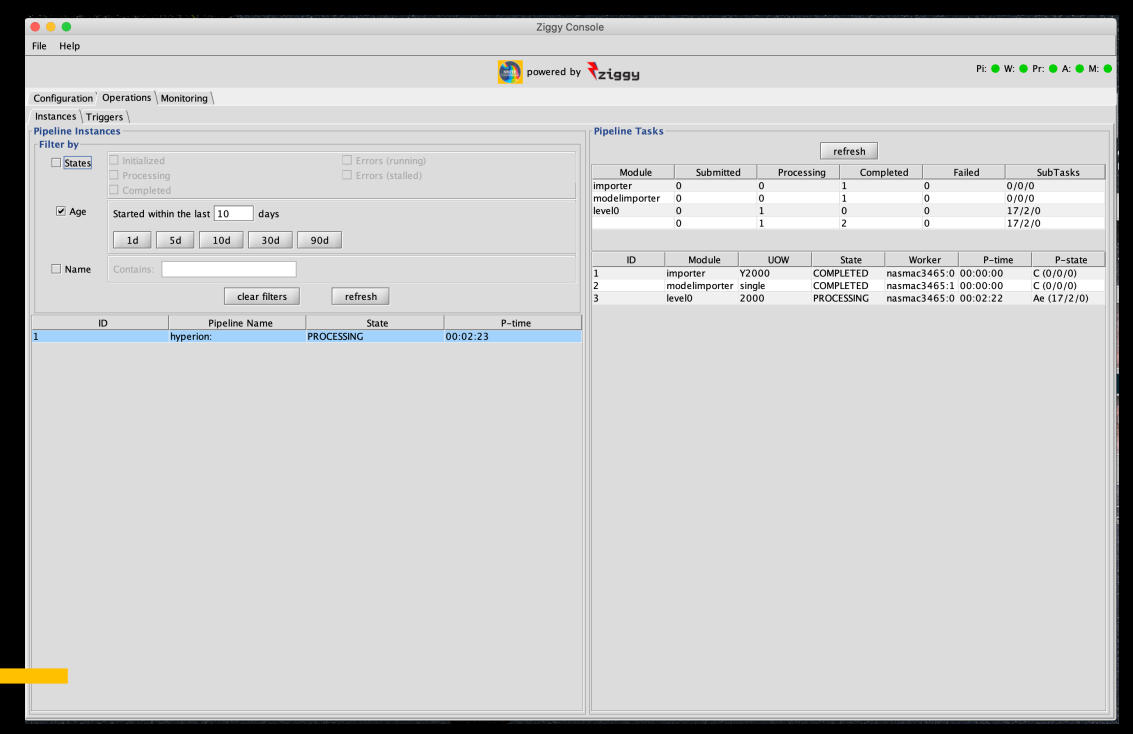
Figure 4. Comparison of surface reflectance spectra retrieved with the Hyperion and AVIRIS sensors, each with ISOFIT and ATREM retrieval algorithms. The results are shown for a vegetation site near Half Moon Bay, CA, observed on April 30, 2015.

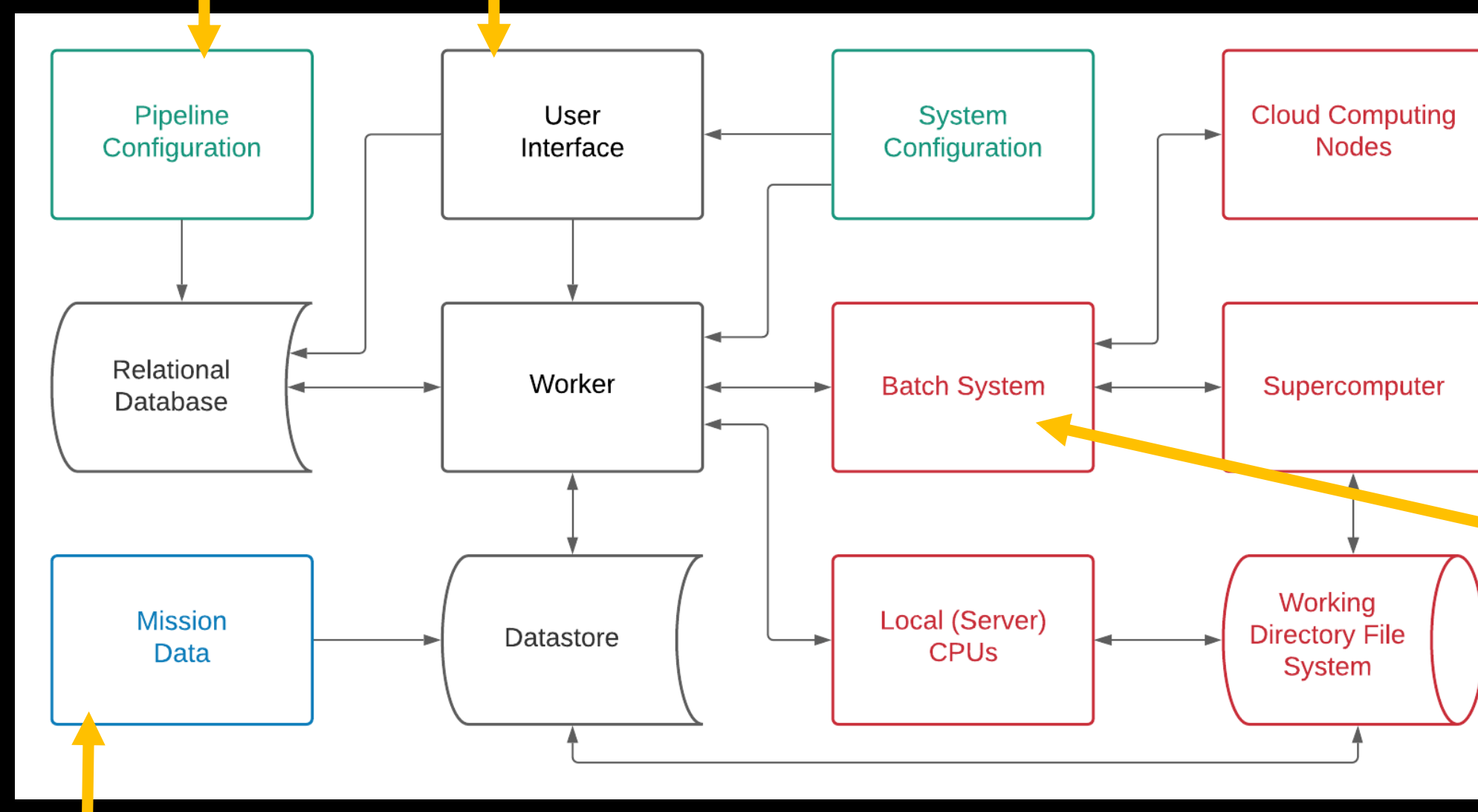
5. Ziggy*, A Portable, Scalable Infrastructure For Science Data Processing Pipelines

*github.com/nasa/ziggy



The mission defines the pipeline: algorithms, execution order, instrument models, etc.





Data, instrument models, etc., can use any desired format. Ziggy supports "keep-up" processing (just process new data) and reprocessing (do everything).

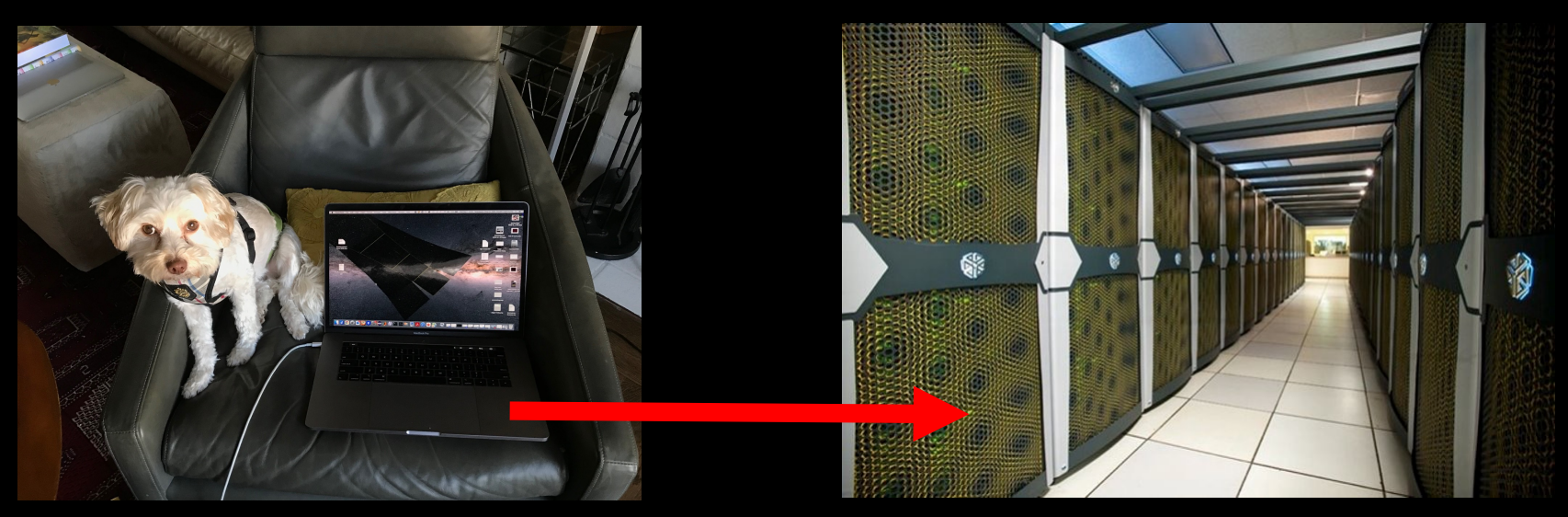
Ziggy supports all scales from local processing on a server to running on petascale systems. Supports hybrid (cloud / HPC) solutions!

Users define a pipeline via a set of XML files that specify the order in which processing algorithms are applied (including optional branching, in which one step is followed by multiple algorithms that run simultaneously), inputs, outputs, and any instrument models or control parameters that are required for each step. Ziggy supports heterogeneous pipelines: each processing algorithm can be in any supported language, and each step can run locally on a server or remotely on a supercomputer or cloud computing facility.

Science data pipelines need to do a lot more than science:

- Logging
- Execution flow
- Execution monitoring
- Data accountability
- Configuration management
- Data marshalling and persistence
- Error handling
- And much more!

Ziggy handles all the Not-Science and lets scientists get on with the Science!



Develop here… … run here!

Ziggy is sufficiently lightweight to run on a laptop and sufficiently robust to run on a supercomputer; builds on Mac OS X and Linux are supported.

6. Processing Data from L0 to L2 on NASA's High-End Computing Capability (HECC) Facility

Large data sets from NASA flight missions can be processed by NASA's Advanced Supercomputer (NAS) Division's supercomputer for as low as $0.115 per Standard Billing Unit (one hour of time on a Broadwell node). HECC enjoys economies of scale comparable to commercial cloud vendors. Projects can purchase their own compute and storage for their exclusive use on margin.

For this estimate we assume a 7-year mission and full reprocessing runs every two years. We take the 55-TB Hyperion data set as a proxy for SBG and scale the 14,000 SBUs needed for processing the Hyperion data set to L2 to 232,432 SBUs/year for SBG. A 64-node Milan architecture can provide 2,130,522 SBUs/year assuming a typical 90% availability.

We include all costs to the project for these resources, including compute, storage, tape backups (2 PB/year for 2 copies of L0), maintenance, operations costs (staff), power, water, facility, software licensing, etc.

69% of the capacity is available for other activities: algorithm development, L3+ processing, more frequent re-processing, *and for supporting open-science activities.*

	Forward (KiloSBU)	Reprocessing (KiloSBU)	Annual (KiloSBU)	Compute Tape	Data Storage (K $)	Projected HECC Cost (K $)
Year						
0				$ 1,719	$ 250	$ 1,969
1	232		232	$ 84		$ 84
2	232		232	$ 84		$ 84
3	232	464	696	$ 84		$ 84
4	232		232	$ 84		$ 84
5	232	928	1160	$ 84		$ 84
6	232		232	$ 84	$ 25	$ 109
7	232	1392	1856	$ 84	$ 25	$ 109
						$ 2,607
					percent utilization	31%

HECC can support forward and reprocessing for SBG from L0 to L2 for approximately $2.6M

[1] "USGS EROS Archive - Earth Observing One (EO-1) - Hyperion," Earth Resources Observation and Science (EROS) Center , 17 April 2019. [Online]. Available: https://www.usgs.gov/centers/eros/science/usgs-eros-archive-earth-observing-one-eo-1-hyperion. [Accessed 1 June 2022].

[2] D. R. Thompson, V. Natraj, R. O. Green, M. C. Helmlinger, B.-C. Gao and M. L. Eastwood, "Optimal estimation for imaging spectrometer atmospheric correction," Remote Sensing of Environment, vol. 216, pp. 355-373, 2018.

[3] B.-C. Gao and A. F. H. Goetz, "Column atmospheric water vapor and vegetation liquid water retrievals from Airborne Imaging Spectrometer data," *J. Geophys. Res.,* vol. 95(D4), p. 3549–3564, 1990.

[4] M. Bouvet, K. Thome, B. Berthelot, A. Bialek, J. Czapla-Myers, N. Fox, P. Goryl, P. Henry, L. Ma, S. Marcq, A. Meygret, B. Wenny and E. Woolliams, "RadCalNet: A Radiometric Calibration Network for Earth Observing Imagers Operating in the Visible to Shortwave Infrared Spectral Range," Remote Sensing, vol. 11, p. 2401, 2019.

[5] B. N. Holben, T. F. Eck, I. Slutsker, D. Tanré, J. P. Buis, A. Setzer, E. Vermote, J. A. Reagan, Y. J. Kaufman, T. Nakajima, F. Lavenu, I. Jankowiak and A. Smirnov, "AERONET - A federated instrument network and data archive for aerosol characterization," Remote Sensing of Environment, vol. 66, pp. 1-16, 1998.

[6] X. Jing, L. Leigh, D. Helder, C. Teixeira Pinto and D. Aaron, "Lifetime Absolute Calibration of the EO-1 Hyperion Sensor and its Validation," *IEEE Trans. on Geoscience and Remote Sensing,* vol. 57, pp. 9466-9475, 2019.