ESSP Executive Committee

The ESSP executive committee will coordinate data challenge efforts related to stellar signal mitigation across the community. With this combined effort, we will insure:

- A centralized way of interacting with the larger community
- Coordinated efforts that minimize repeated effort
- Self-consistent data processing and results across challenges

We will also begin the work to host relevant useful resources, such as:

- A regularly updated list of different methods
- A repository of data from challenges with relevant summary descriptions (i.e. SNR, known systematics, etc.)
- Tutorials

If you have suggestions, feel free to reach out to any of the members.

Current EC Members: Khaled Al Moulla, Michaël Cretignier, João Faria, Jacob Luhn, Federica Rescigno, Lily Zhao

Current lead e-mail: lilylingzhao@uchicago.edu

Intent of New Comparison Work

With ESSP II, we had a difficult time capturing what methods were doing. We moved to using solar data from different instruments to give us a better grasp on the different sources of variation:

1. Planet variation, removed and injected by our will
2. Instrumental variation (backed by four instruments + hopefully some clever modeling)
3. Solar variation (traced by various incomplete measures including resolved images)

For this next step, we will be releasing various forms of the combined solar data (ESSP III) to gauge how methods perform relevant to each of the above sources of variation.

These slides describe

1. The data products and types of data sets that will be shared and
2. The intended analysis to be done with the results.

Specific questions for participants that we would appreciate feedback on are highlighted throughout in **orange**. Answers to these questions or any other comments can also be submitted anonymously through this Google Form or sent to any of the EC members.

Proposed Approximate Timeline

ΔD	Date	Event
	January 22	Share planned analysis (that's this) and ask for feedback (please!)
8	January 30	*Optional Meeting*: (11 AM ET) go through these slides together
4	February 3	**Deadline** for proposed changes to data
7	February 10	Initial data release (or sooner if there are few changes to the proposed data structure)
18	February 28	**Submission**: initial method descriptions and information on contributors
3	week of March 3	*Meeting 1*: questions about the data, pipeline automation, etc.; discuss EPRV 6 abstract
11	March 14	EPRV 6 abstract submission deadline
5	March 19	**Submission**: results with initial data sets
5	week of March 24	*Meeting 2*: discuss initial results, establish possible collaborations all data sets released (unless changes are discussed at meeting 2)
52	May 15	**Final Submission**: final results for all data sets
15	May 30	Share initial analysis of results
10	week of June 9	*Meeting 3*: discuss initial analysis and EPRV 6 presentation
21	June 30	EPRV 6

Is this reasonable? Are there any obvious conflicts?

1. Data Products

Data

Data Products

The data will be standardized s.t. you can treat every file the same regardless of its source instrument.

- Spectra (vacuum BC wavelengths, flux, continuum, blaze, merged 1D)
- Mask of wavelength range shared by all four instruments
- CCFS (order-by-order, combined)
  - Standardized to same velocity grid
- Time series (RVs, indicators, some meta-data and telemetry that is common across all four instruments)

What other data products (e.g. specific instrument telemetry, telluric model, photometry, specific indicators, etc.) does your method absolutely need?

What other standardization of the data might be needed?

A Note on Data Selection

We will be down-selecting the data (i.e. rather than averaging over many observations) to construct time stamps more typical of night-time observing.

The selection procedure emulates getting three shots of a target each “night” with each instrument.

- Given the specific sampling of the combined data set, this equates to ~260 observations per data set
- Though the selection will be random, it will be constrained to a narrow range of high airmass
- Observations will be clustered around some focal time to emulate back-to-back observations
- There will be two kinds of data sets:
  - Present the solar observations as is
  - Combine a small number of observations to integrate over p-modes

Types of Test Sets

The initial idea is to share on the order of ~10 data sets that vary in:

- Injected Doppler shifts (or lack thereof); more details on next slide
- Time sampling
    1. All observations randomly selected
    2. One day of dense sampling (to highlight short-timescale stellar signals)
    3. Observations optimized for overlap between instruments

The data sets will also be standardized with the goal that running different data sets will hopefully approximate:

```
> for dir_name in list_of_data_set_directories:
>     results = method(dir_name)
```

How many data sets is too many data sets? (For reference, ESSP II had 4; Dumusque+ 2017 had 14)

Would you rather (A) run many different data sets in one data challenge, or (B) run fewer, more targeted data sets in each data challenge, but have more data challenges?

It would be most helpful to get ranges (i.e. I would be willing to run up to N data sets per data challenge and no more than M data challenges per year)

Planet Injection

Planet Properties

- The number of planets per system will vary
- Most periods will be constrained by the ~30 day time baseline, and therefore range from 2.5-5 complete phases in 30 days (i.e. 6-12 days)
- Some systems may include a longer period planet
- Relatively low eccentricity
- All planets will be edge-on
- Time of periastrons for multi-planet systems will be randomized

This will be done double blind: the exact parameters for each system will remain hidden.

Are there any other planetary-system-types or detection scenarios that we should include?

Injection Method

1. Generous masking out of telluric regions (but micro-tellurics will likely not be accounted for)
2. Spectral wavelengths shifted
3. Re-derive consequent data products (1D merged spectra, CCF, time series)

- All systems will be checked for
  - Dynamical stability
  - Detectability given cited instrument errors
- PSF changes will not be injected

Standard Planet Recovery Pipeline

In the hopes of standardizing results as much as possible, we will also share an intentionally simple standard planet recovery pipeline. This code will be released for teams to use in tuning methods that do not jointly fit for the planet and stellar signal.

The code will follow the classic script of:

1. Periodogram
2. Fit Keplerian to highest peak
3. Remove best fit planet
4. Iterate until there are no more significant periodogram peaks
5. Refit all planet parameters together

Are there suggested improvements to this pipeline?

2. Intended Analysis

Method Output

Each method will be expected to output and submit the following:

1. "Cleaned" RVs: a time series in which all astrophysical contributions (whether stellar or planetary) have been removed
2. "Activity" RVs: a time series of the modeled out stellar signals
3. Planetary RVs if available
    - For methods that jointly fit for the star and planet, this would be the Keplerian portion of the model
    - For methods that shift and center the data (e.g. before constructing a template, etc.) those shifts would be included in this category
    - For other methods, we will process the cleaned RVs with the specified standard planet recovery pipeline
4. Time series of any "activity indicators" produced by your method

Methods that do not preserve injected planetary signals may be asked to conduct a standardized training/validation analysis.

Evaluating Planet Recovery

Planet detection will be evaluated along similar lines as Dumusque+ 2017. The metrics will include:

- Detection of true planet signals (and corresponding confidence)
- Detection of false positives (and corresponding confidence)
- Nature of false positives (e.g., is the detected period the stellar rotation rate or a harmonic?)
- Accuracy of recovered planet parameters given uncertainties
- Comparison of performance as a function of planet properties, ratio of activity to planet signal, etc.

This May Reveal:

- Are methods successfully preserving injected planet signals?
- What is the expected increase in planet detectability for a given method? Does this change with instrument or activity level/type?
- What is this occurrence rate of false positives? Can they be tied to any solar properties?
- Can differences in planet recovery be tied to performance w.r.t. follow-up training/validation analysis?

Correlation of Activity/Corrected RVs w/ Different Sources of Variation

Sources of variation include:

- Planet
- Stellar
  - SolAster RVs
  - SDO magnetic field and spot properties (coverage, rotating towards/away)
  - Various indicators
  - RVs from other methods
- Instrument
  - Fit model to shared features between instruments (e.g., median filter, GP, other?)

This May Reveal:

- Do some methods trace a certain source of variability (spots, granulation, instrument) better than other sources?
- Can we group methods by what they are tracing/ correlation with each others' results?
- Do some methods perform better/worse on certain instruments?
- Do some methods perform better/worse depending on the nature of the injected planet?

Note: the clarity of this result will depend on (1) how completely we can characterize the sources of variation and (2) how degenerate the correlations may be with each other. Therefore, we will also be characterizing those two factors as a necessary component of this analysis.

RMS Reduction

As with the ESSP II comparison, we can do a straight comparison of reduction in global and intra-night scatter.

Of course, a reduction in RMS is only expected if the true RV time series is assumed to be centered about zero (i.e there is no injected planet signal or the planet signal is successfully preserved and can therefore be subtracted).

Still, it's trivial and possibly interesting to compare the reduction in RMS to other metrics for method performance.

This May Reveal:

- Method efficacy in the case of no injected planets
- How closely does RMS reduction trace other method performance metrics?

Discussion Section Points

- Just one month of data
  - I'll characterize how much variability is in this month vs. solar min/max and expected stellar min/max across exoplanet host stars
  - Discuss how a smaller change in variation can both hurt and help stellar signal models
- Of course all the stuff about how solar observations are necessary but not sufficient for understanding method performance across spectral types
- And many more as we get the results!

Future Directions

- Comparison over a longer time baseline with public data sets (EXPRES (soon), NEID, KPF)
- Dependence of method performance on data quality (e.g. SNR, resolution, wavelength range)
- Analyze the source of any remaining uncorrected signals
- If performance changes with instrument, investigating why
- Test on well-characterized data sets of a range of spectral types

Further discussion planned at an EPRV 6 splinter session led by Khaled and Federica