How good does LIFE have to measure to distinguish Earth at different ages?

Version 0.2

Thomas Birbacher

December 6, 2020

Abstract

One goal of the LIFE space mission is to characterize the atmospheric components of terrestrial exoplanets with a space-based mid-infrared nulling interferometer. A Monte Carlo simulation was used to determine how good this instrument has to measure to confidently distinguish between emission spectra of earth at different epochs and weather conditions, and to evaluate which metric should be used to express how well the spectra can be distinguished.

The three investigated metrics, which are the posterior entropy, the confidence of inference, and the probability of inference not agreeing with the ground truth, all show the same qualitative behaviour, and can all be used for further work.

A confidence of inference of more than 97 % can be achieved with a spectral resolution of 50 and a signal-to-noise ratio of 5.

The source is available at https://github.com/thomabir/LIFE-spectroscopy.

1 Introduction

The goal of the LIFE space mission is to observe a sample of terrestrial exoplanets with a space-based infrared nulling-interferometer, and to characterize their habitability and diversity [2, 4]. One focus of the mission is to find signatures of potential biological activity or at least habitability in the emission spectra of terrestrial planets [3].

This work uses simulated spectra of earth at different epochs and weather conditions [5] to answer two questions:

1. How well does the instrumentation on LIFE have to measure to be able to confidently distinguish these spectra? Which spectral resolution and signal-to-noise ratio are required?

2. What metric should be used to decide how well those spectra can be distinguished, or more generally, how well the experiment is expected to perform?

The earth spectra that should be distinguished are shown in fig. 1.

While there is no reason to expect that the emission spectra of the planets potentially observed by LIFE will be similar to that of earth, this approach constitutes a minimal requirement for the mission: In order for LIFE to be successful, it should at least be able to distinguish earth twins at different stages of evolution. Furthermore, to determine which kind of metric is useful to describe distinguishability, it is easier to start with a small sample of hypotheses to make computations quicker.

Bayesian experimental design [1] is used in section 2 to quantify how well different hypotheses can be distinguished with a given experiment. This mathematical framework associates to each possible outcome of each possible experiment a number, called the utility of the experiment, and provides a way to calculate the utility of an experimental design. In order to use Bayesian experimental design, a utility function has to be chosen, which can be a straightforward one such as the confidence of the inference, or a more complex one such as the entropy of the posterior distribution.

Section 3 describes the setup of a Monte Carlo simulation of the outcomes of many measurements with a hypothetical instrumentation for LIFE. Such a simulation was run for each utility function to find out which one is most suitable to describe how well

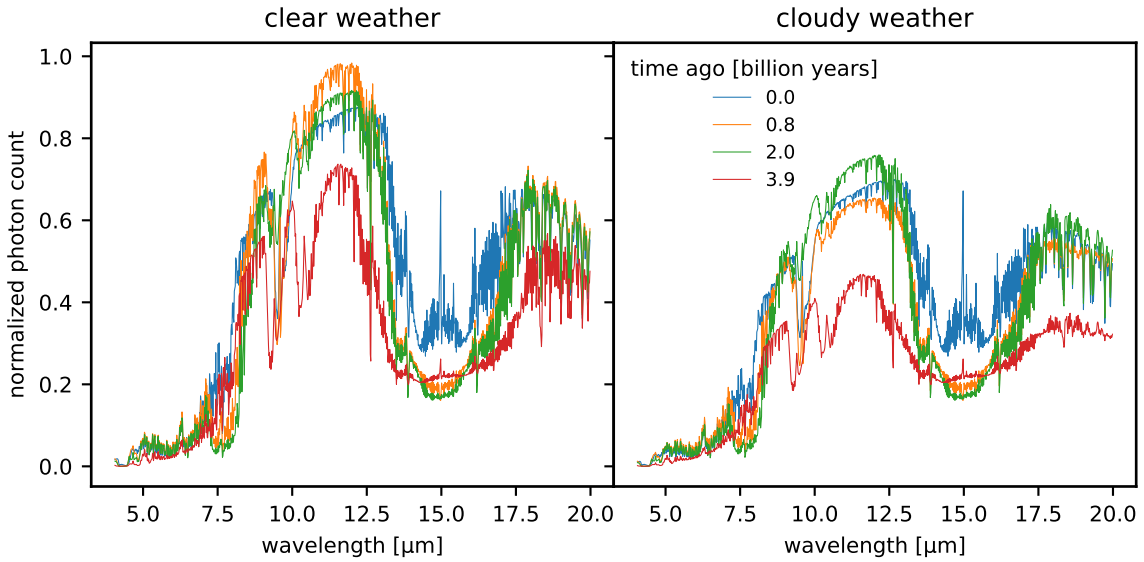


Fig. 1: Simulated infrared spectra of earth at different epochs (modern, 0.8 Gyr ago, 2.0 Gyr ago, 3.9 Gyr ago) and with different weather (clear, cloudy). The goal of this work is to evaluate how good an instrument has to measure to be able to robustly distinguish those spectra.

the different earth spectra can be distinguished.

The results of the simulation in section 4 show the utility of different experimental designs, and in particular the utility of designs with constant exposure time, but varying spectral resolution and signal-to-noise ratio. This illustrates the compromise between these two parameters, and how Bayesian experimental design can be used to find the "ideal" compromise.

2 Theory

The analysis in this report is based on three topics: Bayesian inference, Bayesian experimental design, and information theory.

Bayesian inference is used as a tool for data analysis, specifically to obtain the probability of a hypothesis, given the measured data and the design of the experiment.

Bayesian experimental design is a framework to find an experimental design that is sufficient or even optimal for distinguishing different hypotheses, avoiding false positives, or gaining the maximum possible amount of information. Optionally, constraints and trade-offs between parameters of the experimental design can be imposed.

Finally, information theory formalizes the intuitive notion of information to a mathematical quantity that can be calculated with Bayesian inference.

2.1 Bayesian inference

The following notation is used:

- δ is an experimental design,
- θ is a hypothesis, and
- x is a measured datum, which is a possible realization of a hypothesis and a design.

The probability of measuring x, given the design δ and hypothesis θ, is denoted as $\mathrm{P}(x \mid \delta, \theta)$.

To perform inference about the hypotheses, Bayes' theorem,

$$\mathrm{P}(\theta \mid x, \delta) = \frac{\mathrm{P}(x \mid \theta, \delta)\mathrm{P}(\theta)}{\mathrm{P}(x \mid \delta)},$$

is used to obtain the posterior likelihood $\mathrm{P}(\theta \mid x, \delta)$, which is the probability of hypothesis θ, given the measured datum and the experimental design

2.2 Bayesian experimental design

Bayesian experimental design assigns a utility $U(\delta)$ to every experimental design δ. The goal of the utility is to describe how good the expected performance of an experimental design will be. The task of finding the best possible experiment is then equivalent to maximizing or minimizing the utility, depending on its definition. There are many possible choices for utility functions, such as

1. the expected information gain of an experiment,

2. the expected confidence that the most likely hypothesis is correct, or

3. the probability that the most likely hypothesis agrees with the ground truth.

Since the outcome of a realistic experiment is a realization of a random process with probability distribution $\mathrm{P}(x \mid \delta, \theta)$, we have to calculate the utility $U(x|\delta)$ of every possible realization, and then find the expectation value,

$$U(\delta) = \int \mathrm{P}(x \mid \delta) U(x \mid \delta) \, \mathrm{d}x. \tag{1}$$

2.3 Utility functions

Depending on the goal of the experiment, there are many possible choices for utility functions. Three useful ones are investigated.

2.3.1 *Entropy loss*

A versatile choice of utility is the difference in entropy H between prior $\mathrm{P}(\theta)$ and posterior $\mathrm{P}(\theta \mid x, \delta)$,

$$\begin{aligned} U(x|\delta) &= -H[\mathrm{P}(\theta \mid x, \delta)] + H[\mathrm{P}(\theta)] \\ &= \sum_\theta \mathrm{P}(\theta \mid x, \delta) \log[\mathrm{P}(\theta \mid x, \delta)] \\ &\quad - \sum_\theta \mathrm{P}(\theta) \log[\mathrm{P}(\theta)], \end{aligned}$$

where the definition of entropy in information theory [6] was used:

$$H(X) = -\sum_x \mathrm{P}(x) \log[\mathrm{P}(x)].$$

Since the same flat prior is used throughout the report, the term associated to the prior only shifts the utility by a constant. To make the plots easier to read without having to look back at the definitions, the prior entropy is dropped from the utility, and only the posterior entropy is used:

$$U(x|\delta) = \sum_\theta \mathrm{P}(\theta \mid x, \delta) \log[\mathrm{P}(\theta \mid x, \delta)].$$

The reason why the posterior entropy is a good measure for information is described intuitively below, and is justified in more detail in [6].[1]

[1] When retracing Shannon's original proof that information is entropy, note that Shannon describes the concept of the information of a channel, whereas this work uses it for the information of the posterior distribution, since it is directly available with Bayesian inference. The difference is that the sign of H is flipped, which also becomes evident in the explanation below.

Intuitively, the posterior entropy is a good measure for information, because it is a metric for the "peakedness" of the posterior, which in turn indicates confidence. To see this, compare the best possible with the worst possible experiment:

The ideal experiment rules out all hypotheses except for one, so the posterior is $(0, \ldots, 0, 1, 0, \ldots, 0)$, which has entropy 0. The worst possible experiment leaves us with the posterior just being equal to the prior, which in the unbiased case is $(1/n, \ldots, 1/n)$ for n hypotheses, which has the maximum possible entropy $\log(n)$. Thus, the goal is to design an experiment which minimizes the posterior entropy.

2.3.2 *Confidence of inference*

How confident are we that the most likely hypothesis, given the measured datum and the experimental design, is true?

This utility will be called the confidence of inference, and it is defined as

$$U(x|\delta) = \max_\theta \{\mathrm{P}(\theta \mid x, \delta)\}.$$

This is the maximum of the posterior $\mathrm{P}(\theta \mid x, \delta)$, which is the best candidate that Bayesian inference gives us for the hypothesis. In other words, given the data, the hypothesis is true with a probability of $U(x|\delta)$.

To make comparisons with other utility functions easier, $1 - U(x|\delta)$ is used instead as the utility function in the plots in section 4, such that a low utility corresponds to a useful experiment.

2.3.3 *Probability of inference not agreeing with ground truth*

In a simulation, the true hypothesis from which the data is generated, called ground truth, is directly accessible. Thus, another possibility for the utility is to compare the result of Bayesian inference, which is the hypothesis with maximum posterior likelihood, to the ground truth. This utility is defined as

$$\begin{aligned} &U(x|\delta) \\ &= \begin{cases} 0 & \text{if } \max_\theta\{\mathrm{P}(\theta \mid x, \delta)\} = \text{ground truth} \\ 1 & \text{if } \max_\theta\{\mathrm{P}(\theta \mid x, \delta)\} \neq \text{ground truth.} \end{cases} \end{aligned}$$

The expected value in eq. (1) will then be equal to the probability of conducting an experiment where the result of Bayesian inference will not agree with the ground truth.

The lower the utility, the better the experiment.

3 Method

3.1 Experiment design

The goal of the experiment is to distinguish earth spectra at four different times — modern, 0.8 Gyr ago, 2.0 Gyr ago, and 3.9 Gyr ago — and with both clear and cloudy weather. These eight spectra are the hypotheses that the experiment should be able to distinguish confidently. Specifically, it is assumed that the measured datum will be drawn from a Poisson distribution with the mean given by the hypotheses. This means that there are no nuisance parameters, such as emissions from outside the planet like zodiacal dust, or instrumental noise, and that the only source of uncertainty is the counting noise inherent to a Poisson process.

The experimental design is characterized by the following parameters:

- the observable wavelength range $[\lambda_{\min}, \lambda_{\max}]$,
- the number of wavelength bins n, which are uniformly distributed in the observable wavelength range,
- the spectral resolution SR,
- the peak signal to noise ratio $\mathrm{SNR}_{\mathrm{peak}}$, and
- a quantity t, which is proportional to the exposure time. It is chosen only up to proportionality since the distance to the source, the brightness of the source, the size of the mirrors of the instrument etc. are not exactly known.

For a constant wavelength range, which is utilized in this work, the parameters are related through the relations

$$n \approx 2\,\mathrm{SR}\,\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}},$$
$$t \approx \frac{\mathrm{SNR}_{\mathrm{peak}}^2}{n}.$$

They are only approximations, since there exist spectral resolutions where the corresponding number of bins would be non-integer, such that n has to be rounded.

3.2 Bayesian model

The eight hypotheses are labelled as $\theta = 1, 2, \ldots, 8$; and the spectra used to generate the hypotheses are originally of the form

$$\Lambda^{\theta} = (\lambda_1^{\theta}, \ldots, \lambda_{250000}^{\theta}),$$

where λ_i^{θ} indicates the number of photons per unit time hitting the ith bin of the detector. For each experimental design δ, the spectra have to be resampled so that they have the correct number of bins n. Also, they have to be rescaled such that the desired peak SNR is realized. The spectra after resampling and rescaling are denoted as

$$\Lambda^{\theta,\delta} = (\lambda_1^{\theta,\delta}, \ldots, \lambda_n^{\theta,\delta}).$$

Finally, the random sampling of data is realized as

$$\mathrm{P}(x \mid \theta, \delta) = \prod_{i=1}^{n} \mathrm{Pois}(\lambda_i^{\theta,\delta}, x_i),$$

where

$$\mathrm{Pois}(\lambda, x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

is the probability of drawing x from a Poission distribution with mean λ. For this simulation, 500 samples were drawn for each of the eight hypotheses, yielding $N = 4000$. To find the posterior likelihood, Bayes' theorem,

$$\mathrm{P}(\theta \mid x, \delta) = \frac{\mathrm{P}(x \mid \theta, \delta)\mathrm{P}(\theta)}{\mathrm{P}(x \mid \delta)},$$

is used with a flat prior, and the evidence $\mathrm{P}(x \mid \delta)$ is calculated through marginalization:

$$\mathrm{P}(x \mid \delta) = \sum_{\theta} \mathrm{P}(x \mid \theta, \delta)\mathrm{P}(\theta).$$

3.3 Utility functions

The integral

$$U(\delta) = \int \mathrm{P}(x \mid \delta) U(x \mid \delta)\, \mathrm{d}x$$

is approximated with the Monte Carlo estimator

$$U(\delta) \approx \frac{1}{N} \sum_{x \in X} U(x \mid \delta),$$

where X is a random sample from the distribution $\mathrm{P}(x \mid \delta)$, and N is the number of samples.

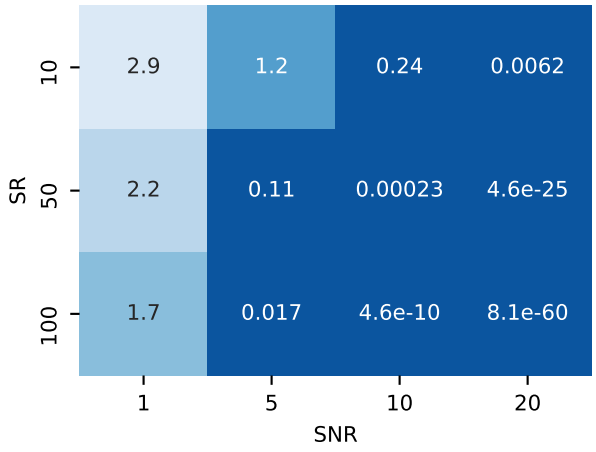
The utility functions in section 2.3 are used.

4 Results & Discussion

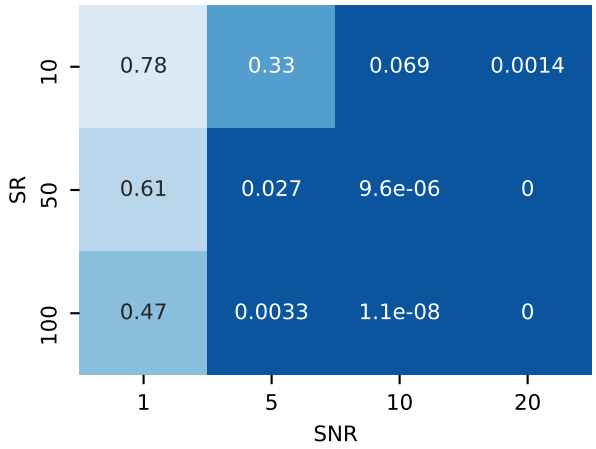
4.1 Heatmaps

To get an overview of the performance of an experiment, a few combinations of SRs and SNRs are compared. The following parameters were used:

(a) Posterior entropy [bits].



(b) 1 − confidence of inference.



(c) Probability of inference not agreeing with ground truth.

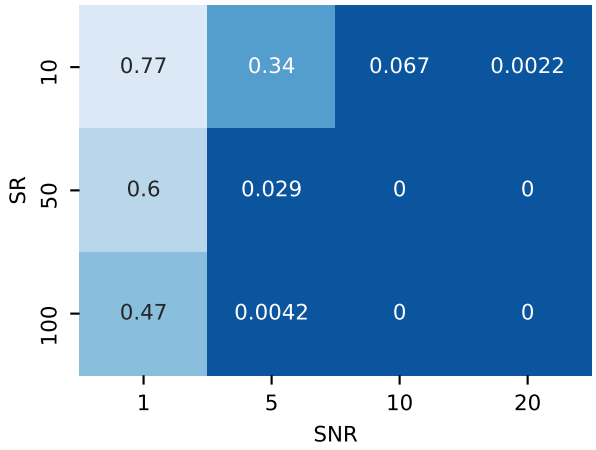


Fig. 2: The utility of experiments with different spectral resolution (SR) and signal-to-noise ratio (SNR). A lower utility indicates a better experiment. As expected, a higher SR and a higher SNR always correspond to a better performing experiment.

- wavelength range: 3 µm to 20 µm,
- SR $\in \{10, 50, 100\}$,
- $\text{SNR}_{\text{peak}} \in \{1, 5, 10, 20\}$,
- number of samples: 4000.

The utilities of experiment designs with different SRs and SNRs are compared in fig. 2. All the utilities are stated in a way such that a lower utility corresponds to a better experiment, which makes comparisons easier.

As intuitively expected, all possible choices for utility functions show that a higher SR and a higher SNR are always rewarded with a better performance of the experiment. A confidence of more than 97 % can be achieved with SR $= 50$ and SNR $= 5$.

In fig. 2b, the numerical precision is sometimes too low to distinguish the utility from zero, since the calculation of the utility usually has to be carried out in linear space rather than in logarithmic space. In fig. 2c, the numerical precision is limited from below by $1/N = 0.00025$. Out of the three utility functions tested, the posterior entropy gives the best numerical precision for a given sample size N.

4.2 Ideal experiments for given exposure times

Given a constant exposure time, SR and SNR are reciprocal to each other, and in practice, a compromise between the two has to be realized in an experiment. Consequently, it is useful to determine which combination of SR and SNR will yield the best performance, given a constant exposure time.

The following experimental parameters were used:

- wavelength range: 4 µm to 20 µm,
- exposure times: $t \in \{400, 1000, 4000\}$
- number of bins: $n \in \{2, 3, 4, \ldots, 200\}$,
- number of samples: 4000.

The utilities of experiments with constant exposure time for distinguishing both clear and cloudy weather are shown in fig. 3. A lower utility corresponds to a better experiment, such that a minimum in the plot indicates the ideal experiment.

Conveniently, the qualitative behaviour of the utility is the same, independent of the choice of utility function, even though their interpretations are different.

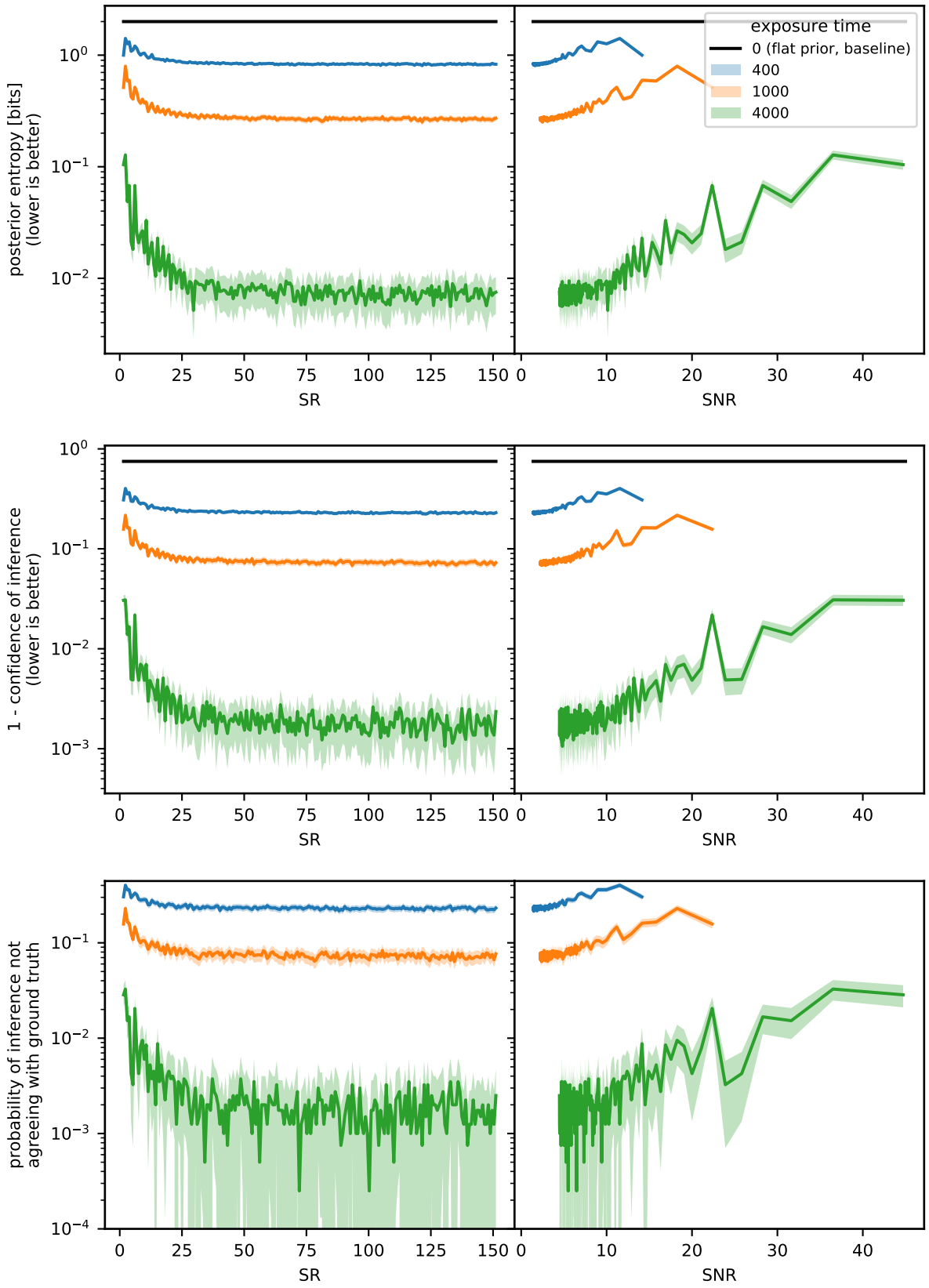


Fig. 3: The utility of experiments with constant exposure times, but different compromises between SR and SNR, for both clear and cloudy weather. The shaded areas are 1σ confidence intervals. All utility functions agree that a higher exposure time results in a better experiment, and that the ideal experiment takes place at a spectral resolution higher than ≈ 30, but that even higher resolutions are not more useful. The data points are inherently discrete, since only an integer number of wavelength bins can be realized, but continuous lines are drawn between the data points as a guide to the eye, so that the fluctuations are clearly visible.

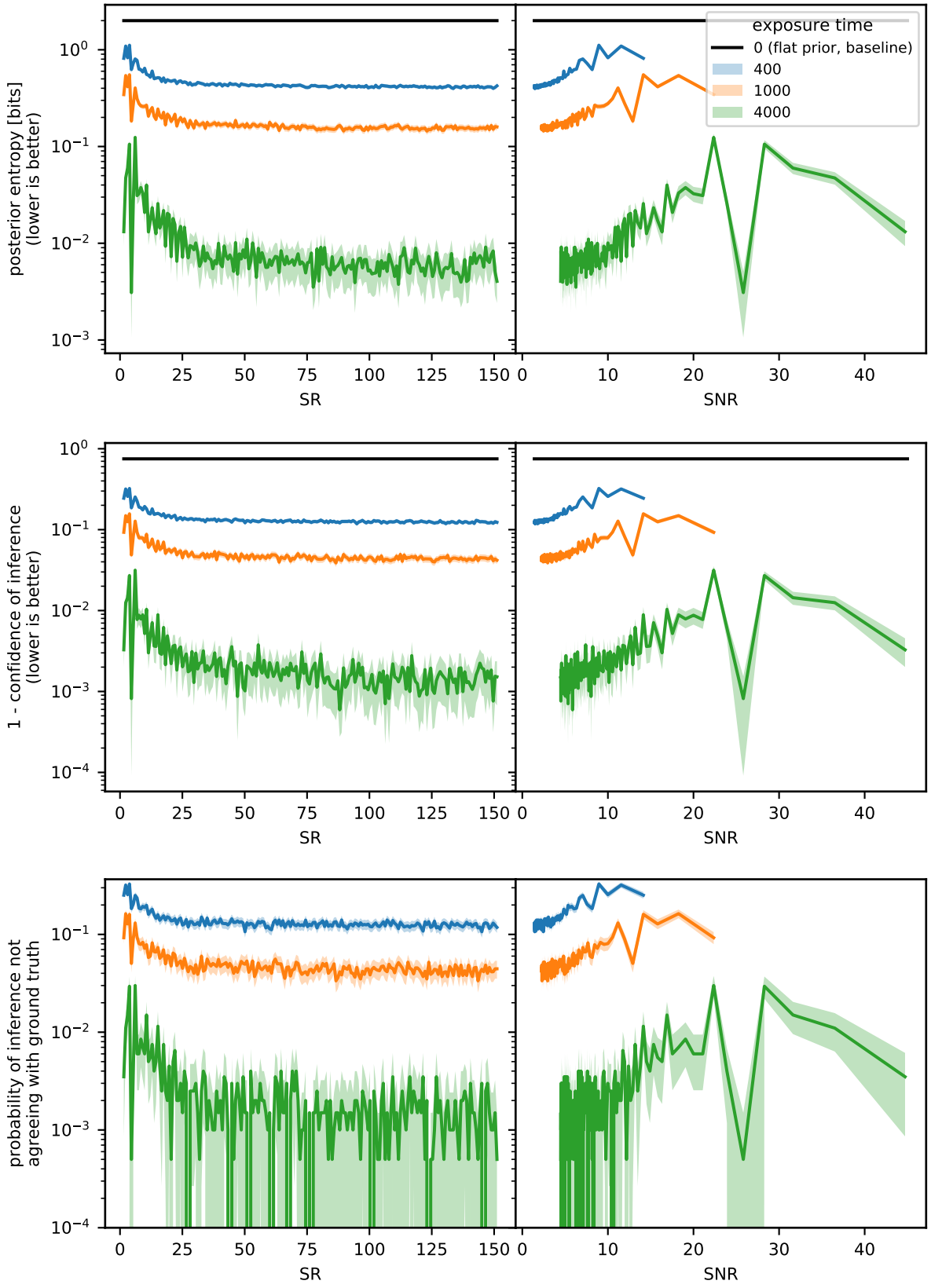


Fig. 4: The utility of experiments with constant exposure times, but different compromises between SR and SNR, for clear weather only. The shaded areas are 1σ confidence intervals. In clear weather conditions, a spectral resolution of 5 is unexpectedly very useful compared to other resolutions. Such an advantage only exists with clear weather, but not when both clear and cloudy atmospheres should be distinguished, as shown in fig. 3.

As expected, all utility functions reward a longer exposure time with better performance. The optimum performance is achieved with a spectral resolution of ≈ 30, but higher resolutions are not beneficial.

The same simulation was also conducted with only the four clear weather spectra in fig. 4, to demonstrate that the resulting ideal experiments depend strongly on the set of spectra that should be distinguished. All utility functions show that an experiment with a spectral resolution of 5 is preferable over most higher resolutions, whereas experiments at SR $= 4$ or SR $= 6$ are considerably worse. This unexpected behaviour shows that in some particular situations, the compromise that usually exists between SNR and SR can be circumvented by choosing the wavelength bins in a way that highlights features where the spectra differ significantly, without requiring a higher resolution. As can be seen by comparing fig. 3 and fig. 4, this advantage at SR $= 5$ disappears when the spectra of cloudy weather are added to the simulation.

5 Conclusion

A Monte Carl simulation was conducted to evaluate how good the infrared spectrometer of the LIFE space mission would have to measure to distinguish the emission spectrum of earth at different times and weathers, and how the robustness and utility of such an experiment can be evaluated.

The three utility functions that were tested are the posterior entropy, the Bayesian confidence of inference, and the probability that the result of Bayesian inference does not agree with the ground truth. All three utility functions show the same qualitative behaviour, but the posterior entropy has the highest numerical precision for a given number of Monte Carlo samples, which gives it an advantage in comparing high-precision measurements.

A confidence of inference of more than 97 % can be achieved with a spectral resolution of 50 and a signal-to-noise ratio of 5. To optimally utilize a given exposure time, the spectral resolution should be 30 or higher.

It should be noted that this analysis only takes into account eight earth spectra in total, whereas the LIFE mission should be able to analyse a much larger variety of planets and atmospheres. Furthermore, an ideal instrument that is always limited by photon noise was assumed, as well as that only the spectrum of the planet will be recorded without any other sources. In order to make a recommendation for the instrumentation of LIFE, a larger variety of sources has to be analysed with a more realistic simulator of the instrument.

References

[1] Kathryn Chaloner and Isabella Verdinelli. "Bayesian experimental design: A review". In: *Statistical Science* (1995), pp. 273–304. URL: https://projecteuclid.org/download/pdf_1/euclid.ss/1177009939.

[2] Jens Kammerer and Sascha P. Quanz. "Simulating the exoplanet yield of a space-based mid-infrared interferometer based on Kepler statistics". In: *Astronomy & Astrophysics* 609, A4 (Jan. 2018), A4. DOI: 10.1051/0004-6361/201731254. arXiv: 1707.06820 [astro-ph.EP].

[3] Sascha P. Quanz et al. *Atmospheric characterization of terrestrial exoplanets in the mid-infrared: biosignatures, habitability & diversity*. 2019. arXiv: 1908.01316 [astro-ph.EP].

[4] Sascha P. Quanz et al. "Exoplanet science with a space-based mid-infrared nulling interferometer". In: *Optical and Infrared Interferometry and Imaging VI*. Vol. 10701. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. July 2018, p. 107011I. DOI: 10.1117/12.2312051. arXiv: 1807.06088 [astro-ph.IM].

[5] S. Rugheimer and L. Kaltenegger. "Spectra of Earth-like Planets through Geological Evolution around FGKM Stars". In: *The Astrophysical Journal* 854.1 (Feb. 2018), p. 19. ISSN: 1538-4357. DOI: 10.3847/1538-4357/aaa47a. URL: http://dx.doi.org/10.3847/1538-4357/aaa47a.

[6] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. URL: https://pure.mpg.de/rest/items/item_2383162_7/component/file_2456978/content.