nature portfolio

https://doi.org/10.1038/s41550-021-01559-x

Supplementary information

Assessing planetary complexity and potential agnostic biosignatures using epsilon machines

In the format provided by the authors and unedited

Supplementary Information for: Assessing Planetary Complexity and Potential Agnostic Biosignatures using Epsilon Machines

Stuart Bartlett[1,2], Jiazheng Li[1], Lixiang Gu[1,3], Lana Sinapayen[4], Siteng Fan[1], Vijay Natraj[5], Jonathan Jiang[5], David Crisp[5] and Yuk Yung[1,5]

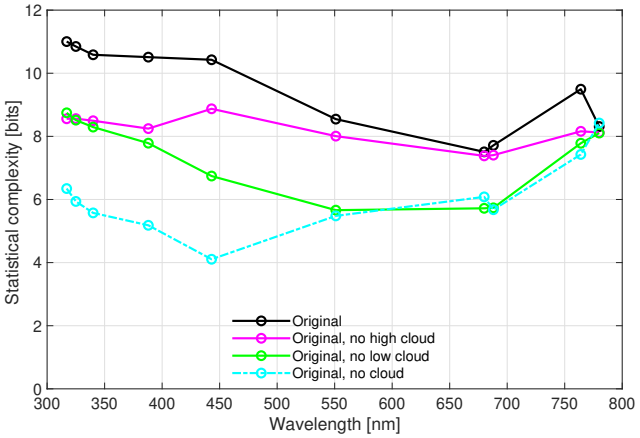
[1*]Division of Geological and Planetary Sciences, California Institute of Technology, 1200 E California Blvd, Pasadena, 91125, California, United States.
[2]Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo, Japan.
[3]Department of Atmospheric and Oceanic Sciences, Peking University, Beijing, 100871, China.
[4]Sony Computer Science Laboratories, Kyoto, Japan.
[5]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, 91109, California, United States.

1 Contributions to Complexity from Planetary Features

In this section we present the statistical complexity values computed for all synthetic Earth types as a function of wavelength. Supplementary Figure 1 illustrates the influence of clouds on statistical complexity, plotted as a function of wavelength. The original (unaltered) data exhibits the highest complexities as expected, followed by the 'no high cloud', 'no low cloud', and cloudless versions. Removing high clouds had little effect on complexity at longer wavelengths (overlap between cyan and green curves). There is an anomalous result at 680nm, where removal of high clouds from the 'no low cloud' version actually increased the complexity. Note that at this wavelength, removing high clouds

from the original time series also caused a negligible decrease in complexity. Hence at these intermediate wavelengths, the majority of the complexity decrease stemmed from the removal of low clouds.

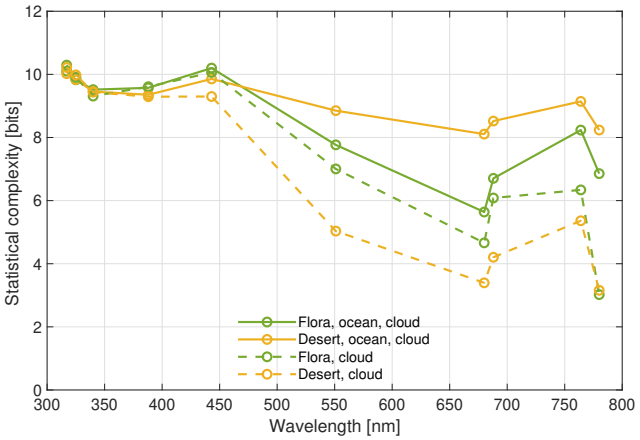
The results at 780nm are ambiguous, where the complexities of all four time series are essentially equal. The distinctions are clearer at lower wavelengths such as 443nm. In the UV we see that removal of high or low clouds causes a similar reduction in complexity, and further cloud removal causes an additional decrease.



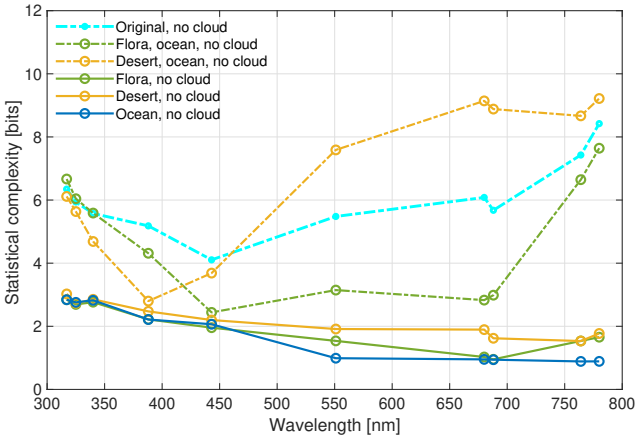
Supplementary Figure 1 Statistical complexity as a function of wavelength for the synthetic Earth time series with modified clouds. The black, red, blue and green curves show values for the original, no high cloud, no low cloud, and cloudless time series, respectively.

We now turn to the influence of the number of surface types for cloudy synthetic Earths. Supplementary Figure 2(a) shows statistical complexity as a function of wavelength for cloudy systems with two surface types (solid green line: vegetation and ocean surface, solid yellow line: desert and ocean surface), and one surface type (dashed green line: pure vegetation surface, dashed yellow line: pure desert surface). We see that in the UV wavelengths, the analysis makes no distinction between the two classes, which is likely due to ozone and Rayleigh scattering obscuring any information from surface features. At longer wavelengths, the number of surface features does have an impact on complexity, due to the lack of absorption in these bands (implying visibility in between clouds), though at 688nm and 764nm, absorption by atmospheric O_2 likely obscures surface information.

Supplementary Figure 2(b) is similar to Supplementary Figure 2(a), but for cloudless synthetic Earths. The cyan dot-dash line corresponds to the original data but with all clouds removed, the green dot-dash line corresponds to only vegetation and ocean, and the yellow dot-dash line corresponds to a desert and ocean world. On average, these cloudless, multi-surface worlds have a higher complexity than the second set, the solid green, yellow and blue lines, corresponding to cloudless flora, desert and ocean worlds, respectively. This



(a) Effect of number of surface types in the presence of clouds



(b) Effect of number of surface types in the absence of clouds

Supplementary Figure 2 Statistical complexity as a function of wavelength for synthetic Earth time series. a) Surface reconstructions in which clouds are left unaltered but surface types are varied, b) Surface reconstructions in which clouds are removed and surface types are varied.

is likely due to the large spatio-temporal contrast in signal between land and ocean for the multi-surface scenarios.

Note that intuitively, one might expect the vegetation-dominated (flora) worlds to exhibit higher complexities than similar abiotic versions such as a desert world. However, when the synthetic Earths are recomposed, the relevant surface image pixels are simply replaced by the respective typical spectra values (such as that for vegetation). Therefore there are no seasonal cycles or other biotic effects in the flora worlds, which are thus no more or less biological than the other synthetic datasets.

Overall, the removal of clouds causes a marked decrease in complexity, in line with expectation. The effects of cloud type changes were more noticeable

at shorter wavelengths. It is also clear that multi-surface worlds have greater complexity than mono-surface worlds, with the discrimination being stronger at longer wavelengths. This indicates a greater contrast between surface types at longer wavelengths. The subtle wavelength-dependent effects are likely due to absorption features (or lack thereof) due to O_2 and ozone. Future work will explain these effects with the help of radiative transfer models.

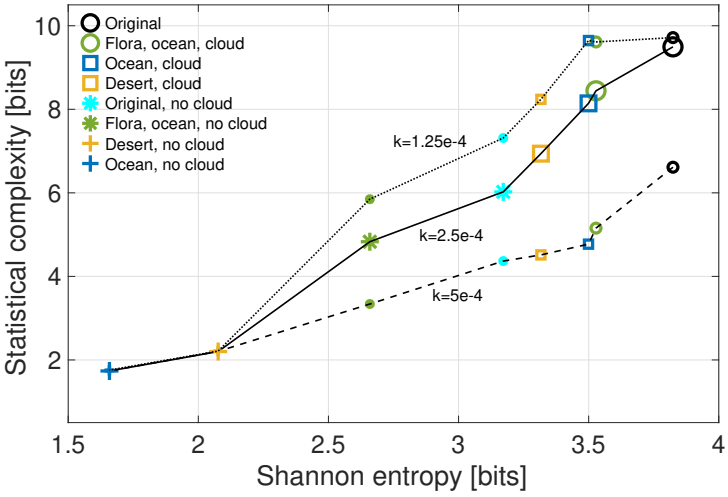
2 Parameter and Data size Sensitivity Analysis

The kernel width parameter k, is an important input to the epsilon machine reconstruction code used in this work [1]. It essentially quantifies the size of the time window over which the algorithm attempts to build a maximally predictive model. Therefore, higher values tend to have a smoothing effect, and eventually, at sufficiently high values of k, the resulting epsilon machine reduces to a single state, single value machine. Thus, excessively high values tend to cause over-smoothing and the algorithm mistakes signal for noise. At the opposite extreme, very small values of k emphasise details at finer timescales. The algorithm may thus mistake noise for data, and make spurious attempts to find patterns in that noise. At excessively low k values the complexity begins to decrease, since the algorithm will revert to simple, purely stochastic epsilon machines as an attempt to match the fine-scale fluctuations.

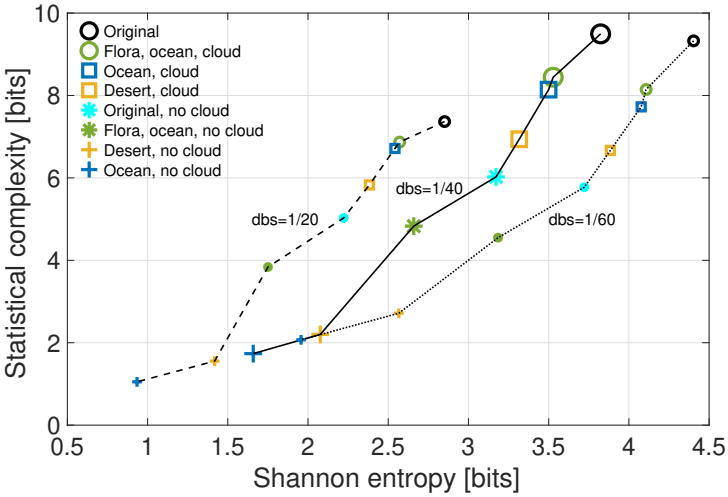
In Supplementary Figure 3(a) we plot the mean complexity and entropy values for a subset of our time series at three different k values. The smaller k value results in higher complexities in general, since the algorithm is attempting to fit finer timescale features in the data. At the higher value of $k = 5e-4$, we see more of a smoothing effect and smaller complexity values. In this case the algorithm is prioritising features at coarser timescales. Despite these variations, it is clear that the general trend shown in Main Text Fig. 5 is unaffected by a doubling or halving of k, and our analysis found that $k = 2.5e-4$ makes the most effective discriminations between the surface types. Note that the simplest synthetic Earths, the cloudless monosurface versions, are essentially unaffected by the choice of kernel width parameter. This is due to the fact that there is little to no structure in those time series for the algorithm to detect, hence changes in k have little impact on complexity.

All input time series also have to be discretised to a finite number of values. Smaller discretisation bin sizes reduce coarse-graining of fine details but incur a higher computational cost. The ideal choice of bin size is a simple trade-off between retaining sufficient detail while allowing feasible computation times. In Supplementary Figure 3(b), the effect of bin size is illustrated for a subset of the time series. As expected, a coarser discretisation (larger bin size) causes a reduction in both entropy and complexity due to the averaging and smoothing effect of larger bins. Using a finer resolution (smaller bin size) tends to increase the entropy but has little effect on complexity. This suggests that our chosen resolution of 40 discrete levels is sufficient to retain the primary features of the time series, since a higher resolution seems to primarily increase the level of

stochasticity. As with varying k, it is clear that discretisation bin size does not impact the general trend illustrated in Main Text Fig. 5. The final phase of our



(a) Effect of kernel width parameter on statistical complexity



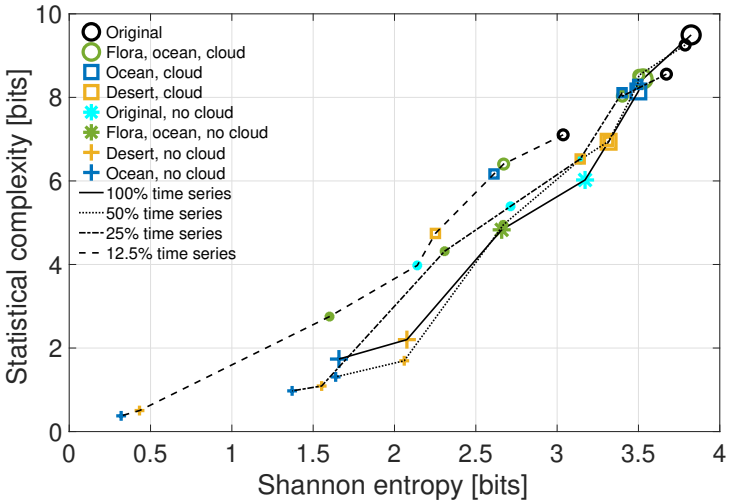
(b) Effect of discretisation bin size on statistical complexity and entropy

Supplementary Figure 3 Statistical complexity as a function of Shannon entropy for three values of a) the kernel width parameter k, and b) the discretisation bin size.

sensitivity analysis compared time series that had been artificially shortened by various fractions. This is a crucial consideration for future applications of our methodology, since long term observations of planetary bodies is technically challenging. We compared time series that had been reduced to a relative

length of 50%, 25%, and 12.5%, and the results are shown in Supplementary Figure 4.

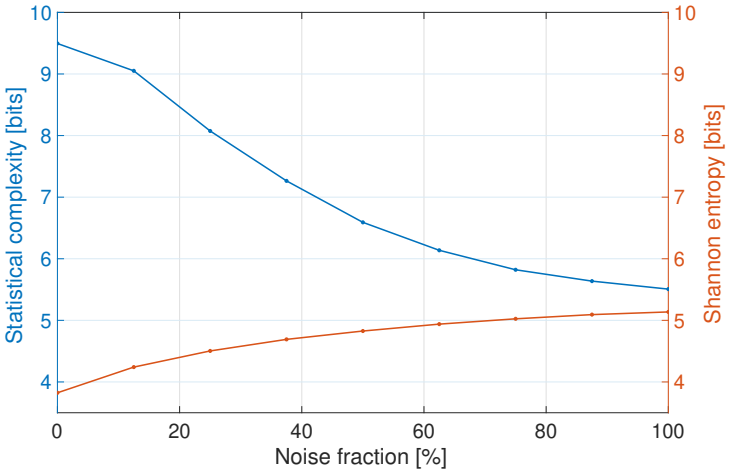
We see that the complexity and entropy values do not show significant changes (reductions) until the data is reduced in length by at least 75%. This suggests that the complexity of these time series is primarily contained within stochastic, rather than reproducible, deterministic or long term features. This can be understood given that the full time series are one year in length, and hence do not contain annual cycles, but only contain diurnal and perhaps other cycles. Since there are so many stochastic features in atmospheric and planetary data such as those used here, most of the structure of the constructed epsilon machines represents an attempt by the algorithm to capture those stochastic features. The ability to find structure in stochastic data (where human eyes might just see noise) is an important ability of EMR.



Supplementary Figure 4 Statistical complexity as a function of Shannon entropy for a subset of synthetic worlds with reduced data lengths.

3 Signal-Noise Sensitivity Analysis

Although our original data contains both intrinsic (natural) and systematic (measurement) noise, it is also important to understand how the time series complexity and entropy values change as the original signal becomes increasingly corrupted. We can readily decrease the signal-to-noise ratio by artificially replacing data points with uniformly distributed (across the unit interval $[0, 1]$) random values. We introduce the continuously-valued noise after normalisation but before discretisation. We performed such a noise sensitivity study on the 'Original' Earth time series and the results are shown in Supplementary Figure 5.



Supplementary Figure 5 Effect of increasing levels of artificial noise on the complexity and entropy of the 'Original' Earth time series.

We see that that increasing noise fractions gradually degrade the signal and the complexity decreases, while the entropy increases. The EMR algorithm is unable to find structure at higher noise levels and hence the resulting epsilon machines reduce to simple random number generators with distributions that converge towards the distribution of the data (this is why the complexity and entropy converge as the noise fraction tends to 100%). It is likely that a slight change (probably an increase) in kernel width parameter would yield exact convergence between complexity and entropy values at the 100% noise level.

4 An Alternative Metric based on the Zip Algorithm

We have presented a complexity analysis based on Epsilon Machine Reconstruction (EMR), but this is one among many such metrics, and measuring complexity is still far from being a settled mathematical formalism. Most complexity metrics tend to be somewhat specific, designed for a certain field or set of applications. In contrast, EMR can be applied to a range of data types, even continuously-varied, multi-variate datasets [2–5] (we are currently employing such frontier techniques to further advance the methodology presented here). One can still ask whether other measures might give similar results to EMR. As mentioned previously, Kolmogorov complexity is essentially a measure of randomness and hence compressibility, but does not have a universal method of computation. A much more common tool is the zip algorithm, used countless times every day for reducing the sizes of files without information loss. This robust algorithm can be used to measure how repetitive or compressible a volume of data is. In the realm of computer science, this notion of minimum description length is directly associated with complexity, which has led to the

use of the zip algorithm as a metric thereof [6]. In order to compare this alternative approach, we zip-compressed all the raw data files used in our analysis, and compared the resulting file sizes. The results are shown in Supplementary Table 1.

Zip rank	Synthetic Earth type	File size [kB]
1	Original	167
2	Desert, ocean, cloud	165
=3	No high cloud	164
=3	No low cloud	164
=3	Flora, ocean, cloud	164
=6	Desert, ocean, no cloud	163
=6	Ocean, cloud	163
=8	No cloud	162
=8	Flora, cloud	162
9	Desert, cloud	160
10	Flora, ocean, no cloud	159
=11	Desert, no cloud	136
=11	Flora, no cloud	136
12	Ocean, no cloud	128

Supplementary Table 1 Alternative ranking based on zip compressibility. Synthetic Earth types are ranked in order of decreasing zip-compressed file size.

We can compare this ranking to the entropy values shown in Main Text Fig. 5, since both are measures of randomness. Although the zip algorithm ranks are approximately similar to the EMR approach, two thirds of the synthetic Earth types lie within $\sim 0.6\%$ of one another, in terms of zip file size. Hence this approach does not provide a strong numerical discriminator of the surface recompositions.

We can also apply this approach to the Jupiter-Earth comparison. The compressed file sizes are shown in rank order in Supplementary Table 2. As with statistical complexity, Earth ranks higher overall.

Zip rank	Time series	File size [B]
1	Earth 443nm	569
2	Earth 551nm	514
3	Earth 760nm	414
4	Jupiter 450.9nm	408
5	Jupiter 568.2nm	341
6	Jupiter 750.5nm	319

Supplementary Table 2 Alternative Earth-Jupiter comparison based on zip compressibility.

In general, the power of EMR, and the reason we used it as the basis of our approach is that it correctly characterises stochastic noise as random but simple (low statistical complexity), whereas most compressibility-based approaches assign high complexities to such data.

References

[1] Brodu, N. Reconstruction of epsilon-machines in predictive frameworks and decisional states. Adv. Complex Syst. **14**, 761–794 (2011) .

[2] Brodu, N. & Crutchfield, J. P. Discovering causal structure with reproducing-kernel hilbert space ϵ-machines. arXiv (2020) .

[3] Marzen, S. & Crutchfield, J. P. Informational and causal architecture of continuous-time renewal processes. J. Stat. Phys. **168**, 109–127 (2017) .

[4] Sinapayen, L. & Ikegami, T. Online fitting of computational cost to environmental complexity: Predictive coding with the ε-network. Artif. Life Conf. Proc. 14 380–387 (2017) .

[5] Marzen, S. E. & Crutchfield, J. P. Structure and randomness of continuous-time, discrete-event processes. J. Stat. Phys. **169**, 303–315 (2017) .

[6] Avinery, R., Kornreich, M. & Beck, R. Universal and accessible entropy estimation using a compression algorithm. Phys. Rev. Lett. **123**, 178102 (2019) .