nature portfolio

Peer Review File

Multi-station deep learning on GNSS time series detects slow slip events in Cascadia

Web links to the author's journal account have been redacted from the decision letters as indicated to maintain confidentiality

12th May 23

Dear Mr Costantino,

Please allow us to apologise for the delay in sending a decision on your manuscript titled "Slow slip detection with deep learning in multi-station raw geodetic time series validated against tremors in Cascadia". It has now been seen by 2 reviewers, whose comments are appended below. You will see that they find your work of some potential interest. However, they have raised quite substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version that fully addresses these serious concerns.

We hope you will find the reviewers' comments useful as you decide how to proceed. Should additional work allow you to address these criticisms, we would be happy to look at a substantially revised manuscript. If you choose to take up this option, please either highlight all changes in the manuscript text file, or provide a list of the changes to the manuscript with your responses to the reviewers.

Please ensure that the revised manuscript meets the following editorial thresholds:

* Present a robust approach to reliably detect Slow Slip Events from raw GNSS data

* Consider more scenarios in generating the synthetic dataset, with respect to both the tectonic signal and data processing strategies.

* Clearly explain and provide sufficient details about how you determine the parameters and thresholds for your detection of slow slip events to the extent that your findings can be reproducible."

Please bear in mind that we will be reluctant to approach the reviewers again in the absence of substantial revisions.

If the revision process takes significantly longer than three months, we will be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Communications Earth & Environment or published elsewhere in the meantime.

We understand that due to the current global situation, the time required for revision may be longer than usual. We would appreciate it if you could keep us informed about an estimated timescale for resubmission, to facilitate our planning. Of course, if you are unable to estimate, we are happy to accommodate necessary extensions nevertheless.

We are committed to providing a fair and constructive peer-review process. Please do not hesitate to contact us if you wish to discuss the revision in more detail.

Please use the following link to submit your revised manuscript, point-by-point response to the reviewers’ comments with a list of your changes to the manuscript text (which should be in a separate document to any cover letter) and any completed checklist:

[link redacted]

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further. Thank you for the opportunity to review your work.

Best regards,

Teng Wang, PhD
Editorial Board Member
Communications Earth & Environment
0000-0003-3729-0139

Joe Aslin
Senior Editor
Communications Earth & Environment

EDITORIAL POLICIES AND FORMAT

If you decide to resubmit your paper, please ensure that your manuscript complies with our editorial policies and complete and upload the checklist below as a Related Manuscript file type with the revised article:

Editorial Policy Policy requirements (Download the link to your computer as a PDF.)

For your information, you can find some guidance regarding format requirements summarized on the following checklist:(https://www.nature.com/documents/commsj-phys-style-formatting-checklist-article.pdf) and formatting guide (https://www.nature.com/documents/commsj-phys-style-formatting-guide-accept.pdf).

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

Using both synthetic and realistic GNSS series, the authors applied deep learning technique trying to automatically detect SSEs in Cascadia. They also compared their detected SSEs with tremor catalog for validation. I think this is a well-written manuscript and a very good study. I recommend it for publication after minor to moderate modification. I have to say I am pretty familiar with SSEs (or ETS), not deep learning. But I think I learned quite a lot about it from reviewing this paper.

In the abstract, “Here we present the FIRST multi-station”. Be careful when you use “first” in a scientific paper. A quick google scholar search using the keep words of “deep learning slow slip detection” indicates several previous studies.

When you generated ultra-realistic synthetic GNSS series, how did you deal with seasonal variation. Sometimes, seasonal change is hardly separated from SSEs.

When I read the manuscript, I found several places where the author used vague or indefinite terms without a clear description. For example, legend of Fig 1, “a HIGHER weight”, how high? Line 94, “we SCALED the amplitudes”, how to scale? And other places.

Line 190-192. When you used 0.4 as the threshold, I am curious how many additional events that were not reported by Michel were detected.

Related to the previous comment, 0.5 or 0.4 confidence value were subjectively chosen by people. I am wondering if you can develop a strategy to objectively determine the confidence value based on some scientific evidence.

Reviewer #2 (Remarks to the Author):

First, please find attached to this review a supplementary material containing the exact same review written below but with the equations appearing correctly.

This manuscript is well written, although I found it difficult to follow the methodologic aspect of this study. While a significant amount of the method is explained in the “Method” section, I feel that the “SSEgenerator”/”SSEdetector” sub-sections (within the “Result” section) were not very clear on the global strategy chosen by the authors. Multiple back and forth between those sections are needed to understand, to first order, what is going on. I think there is room for improvement so that the readers can follow the methodology smoothly (see comments below).

This study aims to build two tools: (1) a generator of synthetic GNSS data which can include the deformation due to synthetic Slow Slip Events (SSEs) occurring on the Cascadia subduction zone (named SSEgenerator); (2) a detector of SSEs using Machine Learning (ML) techniques (named SSEdetector). Their global strategy is to teach the ML algorithm to detect a SSE over a limited temporal window, here 60 days, and to apply this detection procedure on real data as a moving window over time. For the learning process of the algorithm, they create 60,000 samples of synthetic position time series of the size of the temporal window (using SSEgenerator), half of them containing a synthetic SSE positioned at the center of the window, which they feed to the tailored ML algorithm. The procedure allows to evaluate the probability of whether or not a SSE is occurring within this time period, but does not allow to determine its position on the fault. The detection is also linked to the specific setting of the GNSS network in Cascadia. The detection is then applied at each time step of the data and provide the probability of occurrence of a SSE along time. The authors first apply this strategy on synthetic data to evaluate the detection capacity of their method. They then apply it to the real GNSS data and compare their results with the SSE catalog from Michel et al (2019), who used a different signal processing technique, and with the rate of tremors. They show that their methodology is in relatively good agreement with the results from Michel et al (2019) and that they can potentially detect additional events. They also evaluate and compare the duration of their detected events with the estimates from Michel et al. (2019) and those based on tremor activity, which seems to first order also in good agreement.

I find this study interesting and worthwhile publishing. Nevertheless, a few points need to be answered first. Note that I am not an expert in Machine Learning and cannot evaluate the quality on that aspect.

(1) As mentioned before, I found the methodology (whether in section SSEgenerator, SSEdetector, or in the Method section) difficult to follow. A paragraph at the end of the Introduction or at the beginning of the Result section, explaining the overall strategy (similarly to what I have done in the former paragraph) is needed. That would greatly help the reader. Additional points that were difficult to understand are indicated in the “Additional comments” section below.

(2) I found a few typos/problems that might potentially impact, to some degree, their procedure. Equation (2) of the average slip (P15 l.437) seems wrong. If I am correct, there should be a length parameter included in the equation. This impacts the amplitude of the synthetic SSE explored. Additionally the authors could directly use the equation for a rectangle rupture instead of passing through the circular crack version. The authors will find further details in the “Additional comments” section below. Equation (6) seems also incorrect, at least in the spirit of what I think they actually want to do. In their manuscript, the authors use the threshold γ as an actual displacement and not as the percentage of the total displacement. In doing so, the authors actually build synthetic SSEs that are at maximum ~18 days long instead of the 30 days intended.

(3) In Cascadia, SSEs are propagating as pulses and have heterogeneous slip along strike (Schmidt and Gao, 2010; Michel et al., 2019). The authors build instead synthetic SSEs that are not propagating and have homogeneous slip along-strike. The authors should provide a synthetic test of a propagating SSE with homogeneous/heterogeneous final slip to see how SSEdetector fairs. I think there will be no problem since SSEdetector evaluates if there is any SSE at time t, and a propagating SSE can be viewed as the combination of multiple SSEs put together along-strike. Nevertheless, it’s better to have an example to understand how the algorithm proceeds. Besides, the authors train their ML algorithm with events lasting less than 18 days (see comment (2)) while they measure events with a duration over 80 days (Fig4.b). I suspect that what the algorithm is detecting are slices of propagating SSEs and the slices size are dependent on the rise time, i.e. the time for which a fault portion slips during the SSE which is not the total duration of the event since it propagates. If it’s below 18 days, it’s fine.

(4) The authors evaluate the duration of the detected SSEs and state that they detect the start of most SSEs before the increase of tremor burst. The study lacks an analysis of the potential temporal smoothing effect of their methodology. For example, I presume that the probability of occurrence of a SSE will start rising when the real SSE is positioned to right side of the temporal window. There might then be a temporal smoothing effect that will extend the duration estimates. On the opposite, the probability detection threshold of 0.5 will tend to shorten the duration. Can the authors characterize the smoothing effect using their synthetics and use it to correct their duration estimates? This will of course depend on the probability detection threshold.

(5) The full GNSS detrended time series are used to create the surrogate data. The authors should

provide more detail on how they detrend. With just a linear function? Or do they model seasonal and post-seismic processes, and remove earthquake and instrumental offsets? In any case, the authors use time series that contains SSEs, some with very large amplitude and long duration, to build their "realistic" noise. To first order, it's fine. But I wonder if they can improve their detection resolution by using their strategy iteratively. After their first detection, the authors could remove the data during those SSEs and use the remaining data to characterize their noise.

(6) In the synthetics, the synthetic SSEs do not take into account any loading (a steady state slope), while this loading source of deformation will be in the real data. Is it negligible over a 60 days window in terms of detection?

(7) The tremor band seems to go deeper that 60 km depth in southern Cascadia, while the authors limit their detection to SSEs down to 60 km depth. Why not extend deeper?

(8) The ML algorithm uses in the end specific features to detect SSEs. Can the authors retrieve what are the important features to detect SSEs?

While it's important to answer those comments, I do not think that the modifications they require will change drastically the main results of this study. In this respect, I suggest a major revision to allow the authors some time to provide the additional content. You'll find below additional comments. Some of them will detail the main comments above.

Additional comments:

(P2) l.2: Ref for the duration of SSEs (e.g. Behr and Burgmann 2021).

(P2) l.23: Tremors and LFE are not explained before. A sentence mentioning that they are micro-seismicity which are often occurring during SSEs in certain subduction zones is sufficient.

(P2) l.24: What do you mean by "in-situ" geophysical monitoring? It's not clear.

(P2) l.31: "tackle" instead of "tackled"?

(P2 and P3) l.39 and l.48: "end-to-end"? What do you mean by "end-to-end"?

(P3) l.52: I suggest removing the word "systematically". There is always a possibility that the method do not detect small SSEs.

(P3) l.66: It's not directly clear what is done here. It is only after reading the method section that the reader begins to understand the procedure and make sense of this paragraph. Maybe it's the vocabulary that loses the reader. For example, the word "sets" in "sets of geodetic time series" is too vague and other words like "samples", "segments", or "portions" might better to illustrate de procedure. Similarly I did not understand what "single sample" in l.67 really referred to. By changing the word "set" to "sample" and changing the paragraph accordingly, it might clarify the procedure.

(P3) l.66: Why take a 60-day window? If there are 2 SSEs in 60 days, can the method detect the 2 events?

(P3) l.72: “must”? Is the ratio between time series with and without synthetic SSEs important? Why must it be half?

(P3) l.76: “ultra-realistic”? It’s a bit over the top.

(P3) l.80: Specify each time which sub-section in the “Method” section?

(P4) l.89: The authors use the fault geometry Slab-2. If you change the fault geometry, does it change something to the detection? I am mostly thinking about the southern region. The authors stop their fault at 60 km depth, while tremors go deeper than that (Fig1.b). It’s a bit surprising since, in Cascadia, SSEs are occurring within the tremor band. Maybe it’s a problem of the fault geometry at this location which is more difficult to assess. There are, I think, other studies that provide the geometry of the Cascadia subduction zone, and the authors can assess the uncertainty on the geometry.

(P4) l.91: What’s the depth range? Are potential SSEs near the trench, between 0 and 20 km depth, taken into account? Reading the method section, it seems so, but this small precision needs to be clear.

(P4) l.92: I suggest removing “realistic”.

(P4) l.96: Why not more than 30 days. There are SSEs that last longer than 30 days? I acknowledge though that the SSEs in Cascadia propagate and a point on the fault might not experience a SSE for more than 30 days. Same question for 10 days. That might hinder the detection of smaller events that will have shorter durations.

(P4) l.97: Is the amplitude of the synthetic SSEs slip distribution homogeneous along-strike? Does it matter? There is something unclear for me: the ML algorithm has no information about the precise location of the GNSS stations but it still learned to recognize a SSE if certain stations moved together due to a common SSE source defined by the synthetics. There’s then an information of the connectivity of the GNSS station and their sensitivity to SSEs. Am I correct?

(P5) l.124: I suggest changing the sentence “This is done by assigning a weight to the data, with those weights being learnt from the data itself” to “This is done by assigning a weight to the data, with those weights being learnt from the data itself and the a priori knowledge of the labels, i.e. whether there is a SSE or not.”. Is that correct? Otherwise it feels a bit like magic.

(P5) l.125-127: “identify the precise timing of aseismic deformation transients”? But are they not always position at the center of samples?

(P6) Fig.2: There are events with magnitude below 6 while it is said that the range of magnitude tested is in between 6 and 7. Why so?

(P6) l.147: Refer to Fig2.a.

(P6) l.153 and Fig2b: The magnitude threshold is interesting. I would show again the contour of the tremor band in Fig2b as in Fig1.b so that one can compare it with the threshold distribution.

However, considering this slab geometry, the tremors seem to go deeper than 60 km depth. Why not take deeper synthetic events (e.g. down to 80 km depth)?

(P7) l.170: The detector is smoothed in time and the size of the window potentially controls how much it is smoothed. Does changing the window to 20/80 days change this smoothing? Additionally, it is a detector over the whole Cascadia subduction zone. What if there are 2 SSEs (e.g. one in the north and one in the south) happening at the same time?

(P7) l.182-184 and Fig.2 in supplement: The actual location of the events detected by Michel et al. (2019) are further inland, within the band of tremor activity, than what is shown in Fig.2 in the supplement. The magnitude detection threshold should then a bit higher than 6.5.

(P9) Fig.4.b: What's the black rectangle? It's not mentioned in the caption.

(P9) l.210: The authors can refer to Table 1.

(P9) l.215: The authors mention that they detect events shorter than 10 days but they did not train their detection method with synthetic events sorter than 10 days. Can the authors comment on this?

(P9) l.217-218 and Fig4: Note that Michel et al. (2019) provides minimum and maximum estimates of the duration. In any case, those estimates are biased by the some temporal smoothing that will tend to extend their duration. This comparison should be seen in light to this effect. The durations the authors provide might also be affected by a certain temporal smoothing which is not documented in the manuscript. The authors can take advantage of the synthetics to estimate the smoothing effect, which will also be linked with the detection probability threshold to define a SSE (i.e. 0.5 in this study). What is a good detection probability threshold considering the synthetics? How are their duration estimates compared to their true duration? The smoothing effect might be dependent of an events magnitude and/or on the size of the window. The authors can maybe produce a kernel that corrects the initial duration estimate.

(P9) l.231 and Fig4: The authors comment on the points below the diagonal but not above. They detected events that have long durations (~70-80 days) but that are corresponding to short events in Michel et al. (2019). Can the authors comment on it? Is it due to the detection threshold or that SSEdetector cannot distinguish between 2 events if they occur at the same time?

(P11) l.247: I don't see 11 events in Fig3.b.

(P11) l.263-266: I am a bit concerned about the potential smoothing effects which is of relevance to those statements. Please see comment "(P9) l.217-218 and Fig4" for more details.

(P11) l.283 and Fig6 in supplement: Indicate in the caption which version of SSEdetector is the red and blue curve.

(P11) l.285: It's a bit surprising since there are more GNSS stations and thus potentially a better characterization of SSEs by the data. I guess there is a tradeoff between the number of station and the number of missing data.

(P11) l.287-298: The results for the southern region, in the supplement, would be also nice.

(P11) l.349-352: Doubts remain considering potential smoothing effects. See comment “(P9) l.217-218 and Fig4” for more details.

Method: I am not an expert in Machine learning methods. Consequently I do not really understand all the points made regarding the method. My questions may arise as a non-expert in this field.

(P13) l.366-369: “…applying random transformations…” I do not understand the significance of this sentence. Is it necessary? Or does it imply an underlying bias? It’s not clear to me.

(P14) l.374: What is the longest data gap sequence in a 60-day window?

(P14) l.374-377: State why do you train also on 352 stations. I assume it’s for a comparison that will be in the discussion. It was not totally clear at first read.

(P14) l.400: “experimentally”?

(P14) l.403: Why 70% of the data? Why is it necessary to keep 30% in place? It’s not clear. How do the results change if you change this ratio to a 100%.

(P14) l.407: I would change “…in order not to…” to “…to avoid…”. It appears a few times in the manuscript (e.g. l.415).

(P14) l.410: “…SSEdetector…” instead of “…SSedetector…”?

(P14) l.414-420: Not clear. Up to this line, I understand the authors have built surrogate data of a size similar to X and that 70% of the data gaps were shuffled from their initial positions. Are the authors cutting X_R into contiguous sections of size W_L? I am not sure since 60,000 synthetic samples are generated. Or does it mean that there will be duplicates in the 60,000 samples? If X_R is not cut into sections, I do not follow the logic. I suggest the authors explain at the beginning of the section a broad guideline of what they intend to do to help the reader follow.

(P15) l.415: Since the data gaps where already shuffled X_R, why apply again a circular shift? After reading the whole manuscript, I think I understood. But when I read this sentence it was not obvious at all.

(P15) l.423: I guess the authors are selecting randomly the barycenter of the SSE when they mention its location?

(P15) l.432: Why not smaller than 6? Besides, Fig 2.a has events lower than 6.

(P15) l.433-434: Instead of “fault geometry” or “fault radius” I suggest SSEs geometry and radius.

(P15) l.433-440: There might be a typo in the equation for the average slip as it does not depend on a characteristic length. Stress drop is often approximated using the following equation (Kanamori and Brodsky, 2004; Noda et al., 2013):

$(\Delta\sigma) = C M\_0/\rho^3$,
where M_0 is the moment released by an earthquake, ρ is a characteristic length of the rupture and is taken as $\rho=A^{(1/2)}$, A corresponds to the area of the rupture, and C is a geometrical parameter which depends on the shape of the rupture. According to Noda et al. (2013), C is equal to 2.44 for a circular crack, and C= 2.53, 3.02 and 5.21 for rectangular ruptures with aspect ratio of 1, 4 and 16, respectively. In parallel, $M\_0=\mu Au$, where μ is the shear modulus and u the average slip of the event. Thus
$\Delta\sigma =(C\mu Au)/A^{(3/2)} =(C\mu u)/A^{(1/2)}$,
and
$u = (\Delta\sigma A^{(1/2)})/C\mu$.
The average slip depends on the area contrarily to the equation written in the manuscript, and assuming an aspect ratio with W=L/2, it follows
$u = (\Delta\sigma(WL)^{(1/2)})/C\mu=(\Delta\sigma L)/(\sqrt{2} C\mu)$.
The authors could calculate C for an aspect ratio of 2, or take C between 2.53 and 3.02 (i.e. aspect ratio of 1 and 4). In any case, I do not think the value of C will have a strong impact on results, but the equation of slip should be corrected.

(P15) l.439: The aspect ratio of 2 seems small for SSEs in Cascadia. Michel et al., 2019 implies that $M\_w>6$ SSEs in Cascadia have aspect ratios between ~3 and 13. While there are biases when inverting for slip on the fault, I do not think they will change to first order the fact that SSEs in Cascadia have large aspect ratios. How does the aspect ratio of 2 impact the results? Are actual detected SSEs cut into multiple events? What does it change if the authors choose an aspect ratio of 4?

(P15) l.444: Why a lognormal? Using the values provided by the authors, we obtain a stress-drop pdf as showed in Figure R1. Is that the PDF chosen?

Figure R1: Stress drop PDF using $c\_V=10$ and $(\Delta\sigma) =5\ 10^4$ Pa. => Please see PDF

(P16) l.455-462: The term “steady state value of the signal (i.e. D) ” is a bit confusing since the signal is also steady state after the SSE (the position values are just close to zero).

(P16) l.459: t_max should be linked to a displacement of $D\gamma$, and t_min to $D-D\gamma$ (the opposite of what is said in the manuscript). Note that I have written $D-D\gamma$ instead $D-\gamma$ since I interpreted γ as a percentage of the signal amplitude. This implies that $\beta=2/T \ln ((1-\gamma)/\gamma)$. If I was plotting ds(t) taking the equation of β from the authors and trying to generate a SSE with T=30 days and D=10, the SSE will not have a duration of 30 days. It will be instead around 18 days. By doing the modifications as proposed, the authors will correct this problem. Besides, I suggest to the authors to change/remove $ds(D-D\gamma)$ and $ds(D\gamma)$ at l.460-461 as ds is written as a function of time in equation (5). It’s at first sight a bit confusing even though ds is also a function of D. Also, there is an underlying assumption that the loading of the fault is disregarded since the “steady state” slope is 0.

(P18) l.545 and Figure S15: The actual signification of h_L is confusing when looking at the supplementary Figure 15. The topographic prominence is defined by the difference between the height of the peak and the height of its highest base: p=h1-B2. Thus h_L=h1-0.7 p is a height slightly above B2. In the supplementary Figure 15, h_L is represented as a slightly shorter version of p. Additionally, in this figure, there are two steps with the number “3” and the text/figure could be bigger. Finally, some examples using real data, showing the number of tremors per day and the estimated final width, would be nice.

(P19) l.568: What is p''(t) and n''(t) ?
The authors should give in the supplement the timing and duration of each SSE they detected. They should also provide at some point (e.g. external files) the probability of SSE detection (Fig3.a) and other data necessary for plotting Fig4.b and 5.d.

This manuscript is well written, although I found it difficult to follow the methodologic aspect of this study. While a significant amount of the method is explained in the “Method” section, I feel that the “SSEgenerator”/”SSEdetector” sub-sections (within the “Result” section) were not very clear on the global strategy chosen by the authors. Multiple back and forth between those sections are needed to understand, to first order, what is going on. I think there is room for improvement so that the readers can follow the methodology smoothly (see comments below).

This study aims to build two tools: (1) a generator of synthetic GNSS data which can include the deformation due to synthetic Slow Slip Events (SSEs) occurring on the Cascadia subduction zone (named *SSEgenerator*); (2) a detector of SSEs using Machine Learning (ML) techniques (named *SSEdetector*). Their global strategy is to teach the ML algorithm to detect a SSE over a limited temporal window, here 60 days, and to apply this detection procedure on real data as a moving window over time. For the learning process of the algorithm, they create 60,000 samples of synthetic position time series of the size of the temporal window (using *SSEgenerator)*, half of them containing a synthetic SSE positioned at the center of the window, which they feed to the tailored ML algorithm. The procedure allows to evaluate the probability of whether or not a SSE is occurring within this time period, but does not allow to determine its position on the fault. The detection is also linked to the specific setting of the GNSS network in Cascadia. The detection is then applied at each time step of the data and provide the probability of occurrence of a SSE along time. The authors first apply this strategy on synthetic data to evaluate the detection capacity of their method. They then apply it to the real GNSS data and compare their results with the SSE catalog from Michel et al (2019), who used a different signal processing technique, and with the rate of tremors. They show that their methodology is in relatively good agreement with the results from Michel et al (2019) and that they can potentially detect additional events. They also evaluate and compare the duration of their detected events with the estimates from Michel et al. (2019) and those based on tremor activity, which seems to first order also in good agreement.

I find this study interesting and worthwhile publishing. Nevertheless, a few points need to be answered first. Note that I am not an expert in Machine Learning and cannot evaluate the quality on that aspect.

(1) As mentioned before, I found the methodology (whether in section *SSEgenerator*, *SSEdetector*, or in the *Method* section) difficult to follow. A paragraph at the end of the *Introduction* or at the beginning of the *Result* section, explaining the overall strategy (similarly to what I have done in the former paragraph) is needed. That would greatly help the reader. Additional points that were difficult to understand are indicated in the “Additional comments” section below.

(2) I found a few typos/problems that might potentially impact, to some degree, their procedure. Equation (2) of the average slip (P15 l.437) seems wrong. If I am correct, there should be a length parameter included in the equation. This impacts the amplitude of the synthetic SSE explored. Additionally the authors could directly use the equation for a rectangle rupture instead of passing through the circular crack version. The authors will find further details in the “Additional comments” section below. Equation (6) seems also incorrect, at least in the spirit of what I think they actually want to do. In their manuscript, the authors use the threshold γ as an actual displacement and not as the percentage of the total displacement. In doing so, the authors actually build synthetic SSEs that are at maximum $\sim$18 days long instead of the 30 days intended.

(3) In Cascadia, SSEs are propagating as pulses and have heterogeneous slip along strike (Schmidt and Gao, 2010; Michel et al., 2019). The authors build instead synthetic SSEs that are not propagating and have homogeneous slip along-strike. The authors should provide a synthetic test of a

propagating SSE with homogeneous/heterogeneous final slip to see how *SSEdetector* fairs. I think there will be no problem since *SSEdetector* evaluates if there is any SSE at time *t*, and a propagating SSE can be viewed as the combination of multiple SSEs put together along-strike. Nevertheless, it's better to have an example to understand how the algorithm proceeds. Besides, the authors train their ML algorithm with events lasting less than 18 days (see comment (2)) while they measure events with a duration over 80 days (Fig4.b). I suspect that what the algorithm is detecting are slices of propagating SSEs and the slices size are dependent on the rise time, i.e. the time for which a fault portion slips during the SSE which is not the total duration of the event since it propagates. If it's below 18 days, it's fine.

(4) The authors evaluate the duration of the detected SSEs and state that they detect the start of most SSEs before the increase of tremor burst. The study lacks an analysis of the potential temporal smoothing effect of their methodology. For example, I presume that the probability of occurrence of a SSE will start rising when the real SSE is positioned to right side of the temporal window. There might then be a temporal smoothing effect that will extend the duration estimates. On the opposite, the probability detection threshold of 0.5 will tend to shorten the duration. Can the authors characterize the smoothing effect using their synthetics and use it to correct their duration estimates? This will of course depend on the probability detection threshold.

(5) The full GNSS detrended time series are used to create the surrogate data. The authors should provide more detail on how they detrend. With just a linear function? Or do they model seasonal and post-seismic processes, and remove earthquake and instrumental offsets? In any case, the authors use time series that contains SSEs, some with very large amplitude and long duration, to build their "realistic" noise. To first order, it's fine. But I wonder if they can improve their detection resolution by using their strategy iteratively. After their first detection, the authors could remove the data during those SSEs and use the remaining data to characterize their noise.

(6) In the synthetics, the synthetic SSEs do not take into account any loading (a steady state slope), while this loading source of deformation will be in the real data. Is it negligible over a 60 days window in terms of detection?

(7) The tremor band seems to go deeper that 60 km depth in southern Cascadia, while the authors limit their detection to SSEs down to 60 km depth. Why not extend deeper?

(8) The ML algorithm uses in the end specific features to detect SSEs. Can the authors retrieve what are the important features to detect SSEs?

While it's important to answer those comments, I do not think that the modifications they require will change drastically the main results of this study. In this respect, I suggest a major revision to allow the authors some time to provide the additional content. You'll find below additional comments. Some of them will detail the main comments above.

<u>Additional comments:</u>

(P2) l.2: Ref for the duration of SSEs (e.g. Behr and Burgmann 2021).

(P2) l.23: Tremors and LFE are not explained before. A sentence mentioning that they are micro-seismicity which are often occurring during SSEs in certain subduction zones is sufficient.

(P2) l.24: What do you mean by "in-situ" geophysical monitoring? It's not clear.

(P2) l.31: "tackle" instead of "tackled"?

(P2 and P3) l.39 and l.48: “end-to-end”? What do you mean by end-to-end?

(P3) l.52: I suggest removing the word “systematically”. There is always a possibility that the method do not detect small SSEs.

(P3) l.66: It’s not directly clear what is done here. It is only after reading the method section that the reader begins to understand the procedure and make sense of this paragraph. Maybe it’s the vocabulary that loses the reader. For example, the word “sets” in “sets of geodetic time series” is too vague and other words like “samples“, “segments”, or “portions” might better to illustrate de procedure. Similarly I did not understand what “single sample” in l.67 really referred to. By changing the word “set” to “sample” and changing the paragraph accordingly, it might clarify the procedure.

(P3) l.66: Why take a 60-day window? If there are 2 SSEs in 60 days, can the method detect the 2 events?

(P3) l.72: “must”? Is the ratio between time series with and without synthetic SSEs important? Why must it be half?

(P3) l.76: “ultra-realistic”? It’s a bit over the top.

(P3) l.80: Specify each time which sub-section in the “Method” section?

(P4) l.89: The authors use the fault geometry Slab-2. If you change the fault geometry, does it change something to the detection? I am mostly thinking about the southern region. The authors stop their fault at 60 km depth, while tremors go deeper than that (Fig1.b). It’s a bit surprising since, in Cascadia, SSEs are occurring within the tremor band. Maybe it’s a problem of the fault geometry at this location which is more difficult to assess. There are, I think, other studies that provide the geometry of the Cascadia subduction zone, and the authors can assess the uncertainty on the geometry.

(P4) l.91: What’s the depth range? Are potential SSEs near the trench, between 0 and 20 km depth, taken into account? Reading the method section, it seems so, but this small precision needs to be clear.

(P4) l.92: I suggest removing “realistic”.

(P4) l.96: Why not more than 30 days. There are SSEs that last longer than 30 days? I acknowledge though that the SSEs in Cascadia propagate and a point on the fault might not experience a SSE for more than 30 days. Same question for 10 days. That might hinder the detection of smaller events that will have shorter durations.

(P4) l.97: Is the amplitude of the synthetic SSEs slip distribution homogeneous along-strike? Does it matter? There is something unclear for me: the ML algorithm has no information about the precise location of the GNSS stations but it still learned to recognize a SSE if certain stations moved together due to a common SSE source defined by the synthetics. There’s then an information of the connectivity of the GNSS station and their sensitivity to SSEs. Am I correct?

(P5) l.124: I suggest changing the sentence “This is done by assigning a weight to the data, with those weights being learnt from the data itself” to “This is done by assigning a weight to the data, with those weights being learnt from the data itself and the a priori knowledge of the labels, i.e. whether there is a SSE or not.”. Is that correct? Otherwise it feels a bit like magic.

(P5) l.125-127: “identify the precise timing of aseismic deformation transients”? But are they not always position at the center of samples?

(P6) Fig.2: There are events with magnitude below 6 while it is said that the range of magnitude tested is in between 6 and 7. Why so?

(P6) l.147: Refer to Fig2.a.

(P6) l.153 and Fig2b: The magnitude threshold is interesting. I would show again the contour of the tremor band in Fig2b as in Fig1.b so that one can compare it with the threshold distribution. However, considering this slab geometry, the tremors seem to go deeper than 60 km depth. Why not take deeper synthetic events (e.g. down to 80 km depth)?

(P7) l.170: The detector is smoothed in time and the size of the window potentially controls how much it is smoothed. Does changing the window to 20/80 days change this smoothing? Additionally, it is a detector over the whole Cascadia subduction zone. What if there are 2 SSEs (e.g. one in the north and one in the south) happening at the same time?

(P7) l.182-184 and Fig.2 in supplement: The actual location of the events detected by Michel et al. (2019) are further inland, within the band of tremor activity, than what is shown in Fig.2 in the supplement. The magnitude detection threshold should then a bit higher than 6.5.

(P9) Fig.4.b: What’s the black rectangle? It’s not mentioned in the caption.

(P9) l.210: The authors can refer to Table 1.

(P9) l.215: The authors mention that they detect events shorter than 10 days but they did not train their detection method with synthetic events sorter than 10 days. Can the authors comment on this?

(P9) l.217-218 and Fig4: Note that Michel et al. (2019) provides minimum and maximum estimates of the duration. In any case, those estimates are biased by the some temporal smoothing that will tend to extend their duration. This comparison should be seen in light to this effect. The durations the authors provide might also be affected by a certain temporal smoothing which is not documented in the manuscript. The authors can take advantage of the synthetics to estimate the smoothing effect, which will also be linked with the detection probability threshold to define a SSE (i.e. 0.5 in this study). What is a good detection probability threshold considering the synthetics? How are their duration estimates compared to their true duration? The smoothing effect might be dependent of an events magnitude and/or on the size of the window. The authors can maybe produce a kernel that corrects the initial duration estimate.

(P9) l.231 and Fig4: The authors comment on the points below the diagonal but not above. They detected events that have long durations (~70-80 days) but that are corresponding to short events in Michel et al. (2019). Can the authors comment on it? Is it due to the detection threshold or that *SSEdetector* cannot distinguish between 2 events if they occur at the same time?

(P11) l.247: I don’t see 11 events in Fig3.b.

(P11) l.263-266: I am a bit concerned about the potential smoothing effects which is of relevance to those statements. Please see comment “(P9) l.217-218 and Fig4” for more details.

(P11) l.283 and Fig6 in supplement: Indicate in the caption which version of *SSEdetector* is the red and blue curve.

(P11) l.285: It's a bit surprising since there are more GNSS stations and thus potentially a better characterization of SSEs by the data. I guess there is a tradeoff between the number of station and the number of missing data.

(P11) l.287-298: The results for the southern region, in the supplement, would be also nice.

(P11) l.349-352: Doubts remain considering potential smoothing effects. See comment "(P9) l.217-218 and Fig4" for more details.

<u>Method</u>: I am not an expert in Machine learning methods. Consequently I do not really understand all the points made regarding the method. My questions may arise as a non-expert in this field.

(P13) l.366-369: "…applying random transformations…" I do not understand the significance of this sentence. Is it necessary? Or does it imply an underlying bias? It's not clear to me.

(P14) l.374: What is the longest data gap sequence in a 60-day window?

(P14) l.374-377: State why do you train also on 352 stations. I assume it's for a comparison that will be in the discussion. It was not totally clear at first read.

(P14) l.400: "experimentally"?

(P14) l.403: Why 70% of the data? Why is it necessary to keep 30% in place? It's not clear. How do the results change if you change this ratio to a 100%.

(P14) l.407: I would change "…in order not to…" to "…to avoid…". It appears a few times in the manuscript (e.g. l.415).

(P14) l.410: "…SSEdetector…" instead of "…SSedetector…"?

(P14) l.414-420: Not clear. Up to this line, I understand the authors have built surrogate data of a size similar to $\mathbf{X}$ and that 70% of the data gaps were shuffled from their initial positions. Are the authors cutting $\mathbf{X}_{\boldsymbol{R}}$ into contiguous sections of size W_L? I am not sure since 60,000 synthetic samples are generated. Or does it mean that there will be duplicates in the 60,000 samples? If $\mathbf{X}_{\boldsymbol{R}}$ is not cut into sections, I do not follow the logic. I suggest the authors explain at the beginning of the section a broad guideline of what they intend to do to help the reader follow.

(P15) l.415: Since the data gaps where already shuffled $\mathbf{X}_{\boldsymbol{R}}$, why apply again a circular shift? After reading the whole manuscript, I think I understood. But when I read this sentence it was not obvious at all.

(P15) l.423: I guess the authors are selecting randomly the barycenter of the SSE when they mention its location?

(P15) l.432: Why not smaller than 6? Besides, Fig 2.a has events lower than 6.

(P15) l.433-434: Instead of "fault geometry" or "fault radius" I suggest SSEs geometry and radius.

(P15) l.433-440: There might be a typo in the equation for the average slip as it does not depend on a characteristic length. Stress drop is often approximated using the following equation (Kanamori and Brodsky, 2004; Noda et al., 2013):

$$\overline{\Delta\sigma} = C\frac{M_0}{\rho^3},$$

where M_0 is the moment released by an earthquake, ρ is a characteristic length of the rupture and is taken as $\rho = A^{1/2}$, A corresponding to the area of the rupture, and C a geometrical parameter which depends on the shape of the rupture. According to Noda et al. (2013), C is equal to 2.44 for a circular crack, and $C = 2.53$, 3.02 and 5.21 for rectangular ruptures with aspect ratio of 1, 4 and 16, respectively. In parallel, $M_0 = \mu A\bar{u}$, where μ is the shear modulus and $\bar{u}$ the average slip of the event. Thus

$$\overline{\Delta\sigma} = C\frac{\mu A\bar{u}}{A^{3/2}} = C\frac{\mu\bar{u}}{A^{1/2}},$$

and

$$\bar{u} = \frac{\overline{\Delta\sigma}A^{1/2}}{C\mu}.$$

The average slip depends on the area contrarily to the equation written in the manuscript, and assuming an aspect ratio with $W = L/2$, it follows

$$\bar{u} = \frac{\overline{\Delta\sigma}(WL)^{1/2}}{C\mu} = \frac{\overline{\Delta\sigma}L}{\sqrt{2}C\mu}.$$

The authors could calculate C for an aspect ratio of 2, or take C between 2.53 and 3.02 (i.e. aspect ratio of 1 and 4). In any case, I do not think the value of C will have a strong impact on results, but the equation of slip should be corrected.

(P15) l.439: The aspect ratio of 2 seems small for SSEs in Cascadia. Michel et al., 2019 implies that $M_w > 6$ SSEs in Cascadia have aspect ratios between ~3 and 13. While there are biases when inverting for slip on the fault, I do not think they will change to first order the fact that SSEs in Cascadia have large aspect ratios. How does the aspect ratio of 2 impact the results? Are actual detected SSEs cut into multiple events? What does it change if the authors choose an aspect ratio of 4?

(P15) l.444: Why a lognormal? Using the values provided by the authors, we obtain a stress-drop pdf as showed in Figure R1. Is that the PDF chosen?

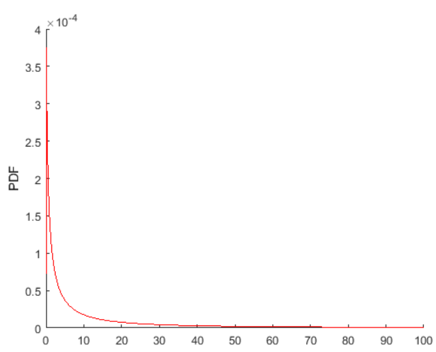


Figure R1: Stress drop PDF using $c_V = 10$ and $\overline{\Delta\sigma} = 5\ 10^4$ Pa.

(P16) l.455-462: The term “steady state value of the signal (i.e. D) ” is a bit confusing since the signal is also steady state after the SSE (the position values are just close to zero).

(P16) l.459: t_{max} should be linked to a displacement of $D\gamma$, and t_{min} to $D - D\gamma$ (the opposite of what is said in the manuscript). Note that I have written $D - D\gamma$ instead $D - \gamma$ since I interpreted γ as a percentage of the signal amplitude. This implies that $\beta = \frac{2}{T}\ln(\frac{1-\gamma}{\gamma})$. If I was plotting $ds(t)$ taking the equation of β from the authors and trying to generate a SSE with $T = 30$ days and $D = 10$, the SSE will not have a duration of 30 days. It will be instead around 18 days. By doing the modifications as proposed, the authors will correct this problem. Besides, I suggest to the authors to change/remove $ds(D - D\gamma)$ and $ds(D\gamma)$ at l.460-461 as ds is written as a function of time in equation (5). It’s at first sight a bit confusing even though ds is also a function of D. Also, there is an underlying assumption that the loading of the fault is disregarded since the “steady state” slope is 0.

(P18) l.545 and Figure S15: The actual signification of h_L is confusing when looking at the supplementary Figure 15. The topographic prominence is defined by the difference between the height of the peak and the height of its highest base: $p = h1 - B2$. Thus $h_L = h1 - 0.7\,p$ is a height slightly above B2. In the supplementary Figure 15, h_L is represented as a slightly shorter version of p. Additionally, in this figure, there are two steps with the number “3” and the text/figure could be bigger. Finally, some examples using real data, showing the number of tremors per day and the estimated final width, would be nice.

(P19) l.568: What is $p''(t)$ and $n''(t)$?

The authors should give in the supplement the timing and duration of each SSE they detected. They should also provide at some point (e.g. external files) the probability of SSE detection (Fig3.a) and other data necessary for plotting Fig4.b and 5.d.

Dear Reviewers,

we would like to thank you for the helpful and constructive comments received on our article. We used all the suggestions to improve the readability of the article and to address all the critical aspects which have been raised.

We agree with reviewer #2 that the paper is, in some parts, difficult to follow. We apologize for this: our work is interdisciplinary and we care that both geoscientists and computer scientists fully understand the content. To improve the readability of the manuscript, we followed reviewer 2's suggestion and wrote a new section, entitled "**Overall strategy**", in which we explain in simple words what are the main goals of the study. We further proofread the language to avoid specific jargon and terminology that could sound vague.
We also added a section to explain the approach used to choose the detection threshold. We compute the ROC (receiver operating characteristic) curve to have a quantitative measure of the impact of the detection threshold on the false positive and true positive rate on synthetic data. Based on these quantitative metrics, we apply the chosen threshold on real data. We made sure to explain all the parameters so that our findings are fully reproducible.

We also added further details about the data processing and how we deal with the tectonic signals. Reviewer 2 had not understood that our analysis is actually performed on "raw" time series, without any post-processing except the Precise Point Positioning (PPP). In the revised version of the manuscript, we better explained that our idea is to keep the GNSS data as it is without any post-processing. We believe that the strength of our approach is that we use non-post-processed GNSS data (we thus keep slow slip event transients as well as post-seismic, common modes, seasonal, etc.), even at the cost of facing a harder problem: this way, no bias is introduced and we can hope to improve the precision. We developed a deep learning method that can learn from data which contain not only the signal of interest (slow slip events), but a variety of other sources of deformation, and we account for all them in our synthetic noise generation strategy.

Reviewer 2 was also concerned about the fact that we train the method on a data set which does not fully cover all the variability and complexity encountered in real data. We thus added two novel sections, that we used to discuss how our method can be generalizable over real data given the choices made for generating synthetic data. During training, deep learning models learn how to approximate a given function linking the inputs and the outputs (in our case, given some input time series, tell if a slow slip event is present or not in the data). The strength of deep learning is that these models approximate the learnt input-output representation by making linear and nonlinear combinations of the features that they learn from the data. It is as if the training data provided some building blocks that can be used (and aggregated) as much as needed by the model to generalize over new, unseen data. To prove that, we added a new section named "**Robustness of SSEdetector to variations in the source characteristics**". In this section, we create four different synthetic case studies to prove that our model is capable, at first order, to:

- detect long (*i.e.,* longer than the longest slow slip event fed during the training phase) and migrating slow slip events, by detecting slices of propagating slow slip events at groups of stations;

- detect short (*i.e.,* shorter than the shortest slow slip event fed during training) slow slip events possibly by learning what is the temporal signature of the signal of interest also at shorter scales;
- detect two slow slip events in the same window, by relying on slices of independent slow slip events
- detect two slow slip events occurring in the same day.

We provide four supplementary figures associated with each test.

Also, we use the 0.5 detection threshold on real data to infer a first-order proxy of the SSE duration. In order to prove the first-order robustness of this procedure, we build a synthetic test case study, which we discuss in a new section of the manuscript "**Validity of the duration proxy against temporal smoothing: the March 2017 slow slip event**" and for which we produced a new figure (Figure 6 of the revised manuscript). We build a synthetic time series by relying on an independent model for the March 2017 slow slip event by Itoh et al. (2022). With this test, we prove that our thresholding procedure compares well with independent results and methodologies.

The Reviewers will find the response to their questions as follows. Questions are colored in black and our answers are reported in blue. Each modification to the text of the manuscript is shown in green, along with the corresponding lines of the revised version of the manuscript. Each modification in the manuscript has been highlighted in yellow for a faster look-up.

We hope that the current version will convince you, so that the paper can be accepted for publication.

Sincerely,
Giuseppe Costantino on behalf of the co-authors

Reviewer #1 (Remarks to the Author):

Using both synthetic and realistic GNSS series, the authors applied deep learning technique trying to automatically detect SSEs in Cascadia. They also compared their detected SSEs with tremor catalog for validation. I think this is a well-written manuscript and a very good study. I recommend it for publication after minor to moderate modification. I have to say I am pretty familiar with SSEs (or ETS), not deep learning. But I think I learned quite a lot about it from reviewing this paper.

1. In the abstract, "Here we present the FIRST multi-station". Be careful when you use "first" in a scientific paper. A quick google scholar search using the keep words of "deep learning slow slip detection" indicates several previous studies.

We agree with the reviewer that this sentence can be misunderstood. Our method leverages a multi-station approach, which is something that was not published before, as far as we know. However, we agree with the reviewer that we should be careful with the word "first". For this reason, we removed it from the manuscript.

2. When you generated ultra-realistic synthetic GNSS series, how did you deal with seasonal variation. Sometimes, seasonal change is hardly separated from SSEs.

Both reviewers had the same comment, which means that we were not clear in our manuscript about that point. We apologize for that. Also, we wrote "raw GNSS time series" which also can lead to confusion (we should have rather written "non-post-processed GNSS time series). When dealing with GNSS time series analysis, people often use trajectory models to subtract the contribution linked to seasonal variations in the GNSS time series. Yet, this can lead in some cases to a poor noise characterization due to the fact that seasonal variations can have different amplitudes in the whole time series and a signal shape that is not necessarily fully reproduced by a simple sum of sines and cosines. This approach can introduce some spurious transient signals which could be erroneously modeled as slow slip events. Another possibility would be to use independent-component-analysis-based methods to extract the seasonal variations. However, the extracted components might contain some useful signals as well as some noise characteristics which could not be removed (e.g., specific harmonics or spatiotemporal patterns). Hence, the main motivation of this work is that we do not want to model any GNSS-constitutive signals, but use the GNSS time series as they are. Of course, this is a difficult approach, but we think that, without a "safe" pre-filtering, it is the only way to extract (small) slow slip events hidden in the noise. We clarified it in the revised manuscript at lines 37-55: "One possibility is to first pre-process the data to remove undesired signals such as seasonal variations, common modes, co- and post-seismic relaxation signals (via denoising, filtering, trajectory modeling or independent-component-analysis-based inversion), but this is at the cost of possibly corrupting the data. Trajectory models are often used to subtract the contribution due to seasonal variations in the GNSS time series. Yet, this can lead, in some cases, to poor noise characterization since seasonal variations can have different amplitudes in the whole time series and a signal shape that is not necessarily fully reproduced by a simple sum of sine and cosine functions. This approach can thus introduce some spurious transient signals which could be erroneously modeled as slow slip events. Another possibility would be to use independent-component-analysis-based methods to extract the seasonal variations. However, the extracted components might contain some useful signals as well as some noise characteristics which could not be removed (e.g., specific harmonics or spatiotemporal patterns). Hence, the main motivation of this work is that we do not want to model any GNSS-constitutive signals, but use the non-post-processed GNSS time series as they are, by relying on a deep learning model which should be able to learn the noise signature, and therefore separate the noise from the relevant information (here, slow slip events)" and, at lines 518-524: "Only the linear trend is removed from the data, without accounting for seasonal variations, common modes, co-seismic or and post-seismic relaxation signals, which we want to keep to have a more realistic noise representation. We remove the linear trend for each station to avoid biases in our noise generation strategy, since the variability due to the linear trend would be captured as a principal component thus introducing spurious piece-wise linear signals in the realistic noise".

3. When I read the manuscript, I found several places where the author used vague or indefinite terms without a clear description. For example, legend of Fig 1, "a HIGHER weight", how high? Line 94, "we SCALED the amplitudes", how to scale? And other places.

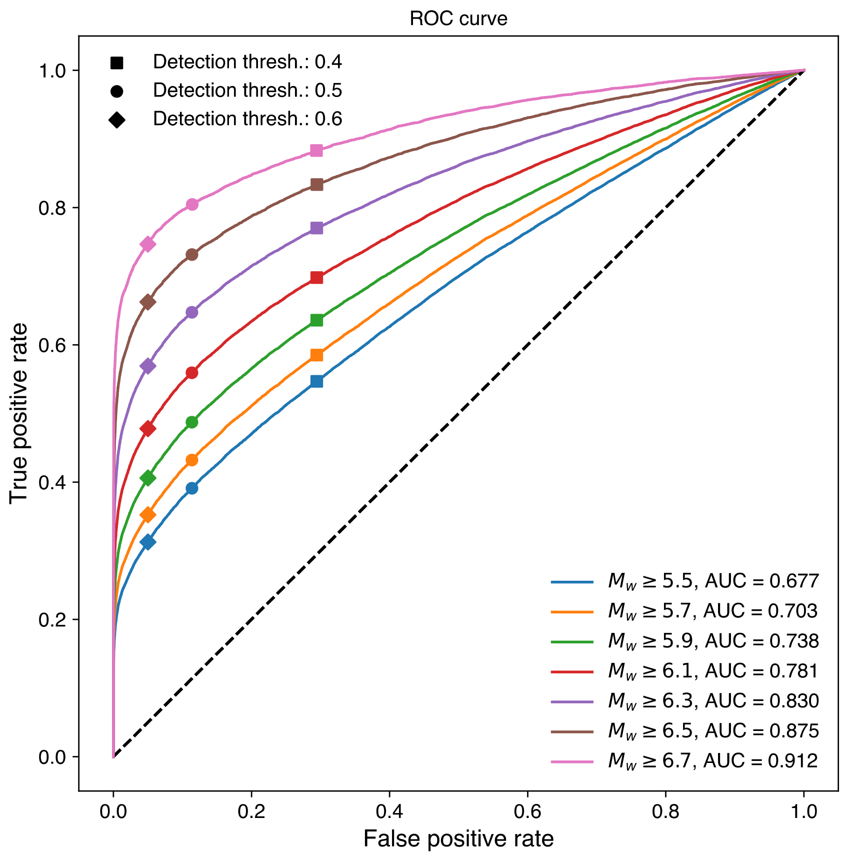
Thank you for the remark. We proofread the manuscript and did our best to clarify the vague and indefinite terms. For example, "We scaled the amplitudes of synthetic SSE signals, modeled as sigmoidal-shaped transients, with a duration following a uniform distribution (from 10 to 30 days)" has been corrected as "We model synthetic SSE signals as sigmoidal-shaped transients, with duration following a uniform distribution (from 10 to 30 days).". We also modified the legend of Figure 1: "the Transformer computes similarities between samples of each station, learning *self-attention* weights to discriminate between the relevant parts of the signals (here, slow slip transients) and the rest (e.g., background noise)"

4. Line 190-192. When you used 0.4 as the threshold, I am curious how many additional events that were not reported by Michel were detected.

When using 0.4 as a threshold, two additional events are detected (total of 37/40). However, also more false positives are introduced (please see also answer to question #5). We added a supplementary figure and clarified it in the paper at lines 259-262: "Interestingly, the two SSEs from Michel et al. with magnitude 5.86 and 6.03, that were missed with a 0.5 probability threshold, can be retrieved when lowering it to 0.4, as shown in Supplementary Figure 18".

5. Related to the previous comment, 0.5 or 0.4 confidence value were subjectively chosen by people. I am wondering if you can develop a strategy to objectively determine the confidence value based on some scientific evidence.

Thank you for this comment. We remade Figure 2 by adding the ROC (receiver operating characteristic) curve as a function of the magnitude ($M_w > M_w^*$), which we use to choose the probability threshold. The reviewer can find the new Figure 2 in the revised manuscript. We also enclose the ROC curve here for more effective look-up.



The ROC curve tells us the discriminative power of a detector as a function of various detection thresholds. It is computed by drawing the true positive rate as a function of the false positive rate for different thresholds. Each curve is obtained by selecting the (positive) events whose magnitude is higher than a set of "limit" magnitudes with increments of 0.2. The negative sample set is the same for all the curves, so that each curve tells us the increment of discriminative SSE power of with respect to the noise. As we see from the figure, the power in detecting SSEs grows as the (SNR) grows, showing that – because of the GNSS noise – small events are hard to retrieve (i.e., they correspond to a lower true positive rate). As said before, each curve is drawn by varying the detection threshold. In our figure, we mark three choices: 0.4, 0.5 and 0.6. As we can see, a threshold of 0.6 is more conservative and leads to fewer false positives, yet at the cost of having less true positives. This translates into retrieving only events with high confidence. Conversely, for a choice of 0.4, the true positives grow, as well as the false positives, since a lower threshold is less conservative. In our problem, the false positive rate is not a big deal, since we use ancillary data (e.g., tremors) to validate our detection and because having a higher true positive rate is preferable to a low true positive rate. We choose to use a threshold of 0.5 to have a good balance between false positives and true positives. With this choice, we tolerate to have more false positives to have more correct detections (with respect to a threshold of 0.6), but, at the same time, we accept to have fewer correct detections without being polluted by a lot of false positives (as it would be in the case of a threshold of 0.4). We added the following in the revised manuscript at lines 192-214:

"First, the probability p output by SSEdetector for each test sample must be classified (i.e., either "noise" or "SSE"). We compute the ROC (receiving operating characteristic) curve as a function of the magnitude, with magnitudes higher than a given threshold, as shown in Figure 2 (a). The curves represent the false positive rate (FPR, the probability of an actual negative being incorrectly classified as positive) as a function of the true positive rate (TPR, the probability that an actual positive will test positive). They are obtained by selecting the (positive) events whose magnitude is higher than a set of “limit” magnitudes with increments of 0.2. The negative sample set is the same for all the curves, so that each curve describes the increment of discriminative SSE power with respect to the noise Each point of the curve is computed with a different detection threshold p_{θ}, such that a sample is classified as SSE when$p > p_{\theta}$. Three threshold choices (0.4, 0.5, 0.6) are marked in Figure 2 (a). As shown in the figure, the higher the magnitude, the higher the number of true positives. When the threshold is high (e.g., 0.6) the model is more conservative: the false positive rate decreases, yet as well as the number of true positives. On the contrary, when the threshold is lower (e.g., 0.4), more events can be correctly detected, at the cost of introducing more false positives. In this study, we choose $p_{\theta} = 0.5$ to have a trade-off between the ability to reveal potentially small SSEs (less false negatives, thus more true positives) and the introduction of false positives. This threshold will be used both for further analyses on synthetic data and for the detection of slow slip events in real data.”.

Reviewer #2 (Remarks to the Author):

This manuscript is well written, although I found it difficult to follow the methodologic aspect of this study. While a significant amount of the method is explained in the "Method" section, I feel that the "SSEgenerator"/"SSEdetector" sub-sections (within the "Result" section) were not very clear on the global strategy chosen by the authors. Multiple back and forth between those sections are needed to understand, to first order, what is going on. I think there is room for improvement so that the readers can follow the methodology smoothly (see comments below).

This study aims to build two tools: (1) a generator of synthetic GNSS data which can include the deformation due to synthetic Slow Slip Events (SSEs) occurring on the Cascadia subduction zone (named SSEgenerator); (2) a detector of SSEs using Machine Learning (ML) techniques (named SSEdetector). Their global strategy is to teach the ML algorithm to detect a SSE over a limited temporal window, here 60 days, and to apply this detection procedure on real data as a moving window over time. For the learning process of the algorithm, they create 60,000 samples of synthetic position time series of the size of the temporal window (using SSEgenerator), half of them containing a synthetic SSE positioned at the center of the window, which they feed to the tailored ML algorithm. The procedure allows to evaluate the probability of whether or not a SSE is occurring within this time period, but does not allow to determine its position on the fault. The detection is also linked to the specific setting of the GNSS network in Cascadia. The detection is then applied at each time step of the data and provide the probability of occurrence of a SSE along time. The authors first apply this strategy on synthetic data to evaluate the detection capacity of their method. They then apply it to the real GNSS data and compare their results with the SSE catalog from Michel et al (2019), who used a different signal processing technique, and with the rate of tremors. They show that their methodology is in relatively good agreement with the results from Michel et al (2019) and that they can potentially detect additional events. They also evaluate and compare the duration of their detected events with the estimates from Michel et al. (2019) and those based on tremor activity, which seems to first order also in good agreement.

I find this study interesting and worthwhile publishing. Nevertheless, a few points need to be answered first. Note that I am not an expert in Machine Learning and cannot evaluate the quality on that aspect.

1. As mentioned before, I found the methodology (whether in section SSEgenerator, SSEdetector, or in the Method section) difficult to follow. A paragraph at the end of the Introduction or at the beginning of the Result section, explaining the overall strategy (similarly to what I have done in the former paragraph) is needed. That would greatly help the reader. Additional points that were difficult to understand are indicated in the "Additional comments" section below.

We are deeply thankful to the reviewer for this comment and for their precise summary. We agree that this section is needed to understand the paper, which can be quite difficult to follow for non-machine-learning specialists. We took inspiration to the reviewer's suggestion to write a new section ("Overall strategy") at the end of the introduction:

"In this study, we build two tools: (1) SSEgenerator, which generates synthetic GNSS time series incorporating both realistic noise and the deformation signal due to synthetic slow slip events along the Cascadia subduction zone, and (2) SSEdetector, which detects slow slip events in GNSS time series by means of deep learning techniques. The overall strategy is to train SSEdetector to reveal the presence of a slow slip event in a fixed-size temporal window, here 60 days, and to apply the detection procedure on real GNSS data with a window sliding over time. We first create 60,000 samples of synthetic GNSS position time series using SSEgenerator, which are then fed to SSEdetector for the training process. Half of the training samples contain only noise, with the remaining half containing an SSE of various sizes and depths along the subduction interface. Thanks to this procedure, SSEdetector can evaluate the probability of whether or not an SSE is occurring in the window, yet it does not allow it to determine its position on the fault. It is important to note that the detection is also linked to the GNSS network in Cascadia. On real data, the detection is applied to each time step and provides the probability of occurrence of a slow slip event over time. We first apply this strategy to synthetic data to evaluate the detection power of the method. Then, we apply SSEdetector on real GNSS time series and compare our results with the SSE catalog from Michel et al. (2019), who used a different signal processing technique, and with the tremor rate.".

2. I found a few typos/problems that might potentially impact, to some degree, their procedure. Equation (2) of the average slip (P15 l.437) seems wrong. If I am correct, there should be a length parameter included in the equation. This impacts the amplitude of the synthetic SSE explored. Additionally the authors could directly use the equation for a rectangle rupture instead of passing through the circular crack version. The authors will find further details in the "Additional comments" section below. Equation (6) seems also incorrect, at least in the spirit of what I think they actually want to do. In their manuscript, the authors use the threshold γ as an actual displacement and not as the percentage of the total displacement. In doing so, the authors actually build synthetic SSEs that are at maximum ~18 days long instead of the 30 days intended.

Thank you for spotting this error. The reviewer is right, a length parameter is missing from equation 2. We corrected the typo in the revised manuscript (it was just a typo since the code of SSEgenerator uses the right formula). Also, the code used for the generation of the synthetic SSE signals (sigmoids, cf. equation 6) is also correct. There is no bias related to this 18-long SSE problem that the reviewer is mentioning, since the equation we are using for SSEgenerator has the same form as the one suggested by the reviewer in his comments (see also minor comments). We agree with the reviewer that this may lead to confusion. Therefore, we followed the reviewer's suggestion and we modified equation 6 in the manuscript as suggested, that is following the codes that we already had developed:

$$\beta = \frac{2}{T}\ln\left(\frac{1}{\gamma} - 1\right)$$

3. In Cascadia, SSEs are propagating as pulses and have heterogeneous slip along strike (Schmidt and Gao, 2010; Michel et al., 2019). The authors build instead synthetic

> SSEs that are not propagating and have homogeneous slip along-strike. The authors should provide a synthetic test of a propagating SSE with homogeneous/heterogeneous final slip to see how SSEdetector fairs. I think there will be no problem since SSEdetector evaluates if there is any SSE at time t, and a propagating SSE can be viewed as the combination of multiple SSEs put together along-strike. Nevertheless, it's better to have an example to understand how the algorithm proceeds. Besides, the authors train their ML algorithm with events lasting less than 18 days (see comment (2)) while they measure events with a duration over 80 days (Fig4.b). I suspect that what the algorithm is detecting are slices of propagating SSEs and the slices size are dependent on the rise time, i.e. the time for which a fault portion slips during the SSE which is not the total duration of the event since it propagates. If it's below 18 days, it's fine.

Thank you for this comment. We find that the reviewer's remark is very relevant and we address it in details in the revised manuscript. The reviewer's comments can be resumed as: "How does the variability of the source parameters affect the prediction of SSEdetector?". In other words, can the authors comment on the fact they generate a very simple synthetic SSE database (*e.g.*, without slow slip propagation, and with a limited variability in terms of stress drop or aspect ratio) and they can retrieve also SSEs that do not meet the characteristics of the SSEs used in the training set? We take this opportunity to make an extensive analysis to assess the robustness of our approach with respect to some source characteristics that are not taken into account in the training database, *e.g.*:

- the aspect ratio and shape of the SSE rupture,
- the length of the synthetic SSE with respect to the fact that SSEdetector is trained on SSEs that last from 10 to 30 days: can SSEdetector retrieve short (< 10 days) SSEs? Can SSEdetector retrieve long (> 60 days) and migrating SSEs?
- the number of SSEs in the same window
- the number of SSEs in the same window occurring at the same time.

We addressed these points by writing a new paragraph entitled "**Robustness of SSEdetector to variations in the source characteristics**" and we wrote the following text:

"We perform an extensive study to test how SSEdetector performs on SSEs that exhibit characteristics in the source parameters that differ from the synthetic data set used for the training. We train SSEdetector on synthetic dislocations with fixed aspect ratios and shapes. However, SSEs in Cascadia generally have aspect ratios between 3 and 13 and exhibit heterogeneous slip along-strike with pulse-like propagation (Schmidt and Gao, 2010; Michel et al., 2019). We create a synthetic test sample containing three slow slip events that propagate in space and time, to simulate a real large SSE that lasts 60 days and propagates southwards, as shown in Supplementary Figure 19. The detection probability and the associated duration proxy suggest that the probability curve can still be used to retrieve SSEs that last longer than 30 days. Thanks to its multi-station approach, our method detects slices of slow slip events through time and space (station dimension), regardless of the training configuration. This can be seen as a stage in which some building blocks are provided to the model, that combines them during the testing phase to detect events having different shapes and sizes. For this reason, SSEdetector can be used to generalize, at first order, over more complex events in real data without the need to build a complex train data set, which should also account for realistic event propagation mechanisms, yet difficult to model.

We also test the generalization ability of SSEdetector over shorter (less than 10 days) events. We build a synthetic sample of a 3-day slow slip event. As shown in Supplementary Figure 20, SSEdetector is able to correctly detect it, with an inferred duration of 4 days, suggesting that it has learned what is the first-order temporal signature of the signals of interest, being able to detect them also at shorter scales.

SSEdetector is applied to real data by following a sliding-window strategy. In real data, two events could be found in the same window, most likely one in the northern and the other in the southern part. For this, we build two synthetic tests, the first test with two 20-day-long events that do not occur at the same time, and the second one with two 20-day-long synchronous events. Supplementary Figures 20 and 21 show the results for both cases, which suggest that SSEdetector can generalize well also when there are two events in the same window, also corroborating the results found on real data.".

4. The authors evaluate the duration of the detected SSEs and state that they detect the start of most SSEs before the increase of tremor burst. The study lacks an analysis of the potential temporal smoothing effect of their methodology. For example, I presume that the probability of occurrence of a SSE will start rising when the real SSE is positioned to right side of the temporal window. There might then be a temporal smoothing effect that will extend the duration estimates. On the opposite, the probability detection threshold of 0.5 will tend to shorten the duration. Can the authors characterize the smoothing effect using their synthetics and use it to correct their duration estimates? This will of course depend on the probability detection threshold.

We thank the reviewer for this interesting comment. To assess a potential smoothing effect on synthetic data, we chose to perform a synthetic test as follows. We compute a noise time series thanks to SSEgenerator. Then, we add it to the surface displacement predicted by the kinematic model of the slow slip event that occurred in March 2017 (Itoh et al, 2022). The duration of the modeled SSE is known (30 days). It is therefore a good case to assess whether SSEgenerator and SSEdetector introduce a temporal smoothing. As shown in the new Figure 6, the duration computed by SSEdetector is in good accordance (28 days) with the one found by Itoh et al., which confirms that our method does not introduce any first-order temporal smoothing. It is worth to remark that the duration is a first-order proxy and should not be overinterpreted (see also answer to question #27, as well as answer to questions #4 and #5 of reviewer #1). We added a paragraph in the revised manuscript entitled "**Validity of the duration proxy against temporal smoothing: the March 2017 slow slip event**" with the following text:

" We test the first-order effectiveness of the duration proxy computation to assess any potential temporal smoothing effects. We focus on the M_w 6.7 slow slip event that occurred on March 2017. We rely both on the measured time series and on the kinematic model of the slip evolution by Itoh et al. (2022). Figure 6(a) shows the displacement field output by Itoh et al.'s model and its temporal evolution is shown in Figure 6(b). We use SSEgenerator to build a random noise time series, to which we add the modeled displacement time series to build a synthetic time series reproducing the March 2017 SSE, as shown in Figure 6(c). Figure 6(e) shows the probability curve output by SSEdetector on the synthetic time series (in blue), to be also compared to the prediction associated with the same event in real GNSS data (in red, see also Figure 3(d)). By the analysis of the probability curve, we estimate a duration of 28 days, comparable with the duration inferred by Itoh et al. (2022) (30 days). A 3-day time shift is found between the probability curve and the slip evolution, represented by moment rate function in Figure 6(d), which can be imputed to the specific realization of synthetic noise. When comparing the probability curve associated with the same event as computed from real data (red curve in 6(e)), we find a better resemblance, consistent also with the catalog from Michel et al. (see Figure 3(d)), which demonstrates that SSEdetector does not suffer from any first-order temporal smoothing issues".

5. The full GNSS detrended time series are used to create the surrogate data. The authors should provide more detail on how they detrend. With just a linear function? Or do they model seasonal and post-seismic processes, and remove earthquake and instrumental offsets? In any case, the authors use time series that contains SSEs, some with very large amplitude and long duration, to build their "realistic" noise. To first order, it's fine. But I wonder if they can improve their detection resolution by using their strategy iteratively. After their first detection, the authors could remove the data during those SSEs and use the remaining data to characterize their noise.

Thank you for the comment. Please first refer to the answer to question #2 by reviewer #1. That said, we detrend by fitting a straight line, without modeling other contributions since we want to build a synthetic noise which resembles the real data at best. About the strategy of iteratively refining the noise: we adopted a conservative approach. This means that on one hand we do not want to model signals like seasonal, post-seismic etc. in GNSS data. On the other hand, this also means that if we followed this iterative "removal" suggested by the reviewer, we could introduce potential bias, as in the aforementioned case. For these reasons, we decided not to "touch" the data and to perform no post-processing on it. We want also to remark that the detrending is not necessary from the point of view of the detection, since deep learning models like CNNs with pooling are invariant to translations and to the linear trend since the convolution operation is inherently performed at a local scale. The only reason why we detrend the data is because a strong trend can negatively influence the principal component analysis, making spurious piecewise linear signals to appear in the time series. For this reason, we detrend the data before passing it to SSEgenerator. We clarified it in the text, please refer to answer to question #2 of reviewer #1.

6. In the synthetics, the synthetic SSEs do not take into account any loading (a steady state slope), while this loading source of deformation will be in the real data. Is it negligible over a 60 days window in terms of detection?

Thank you for the comment. We think that, at first order, this is negligible. Also, it is easy for deep learning methods to estimate a straight line from the data. For this reason, we think that the presence of the interseismic loading in the window is negligible.

7. The tremor band seems to go deeper that 60 km depth in southern Cascadia, while the authors limit their detection to SSEs down to 60 km depth. Why not extend deeper?

Thank you for this comment. Albeit it would be worth going deeper than 60km depth would be interesting to assess the sensitivity of the network, we find that the SNR in these cases would be very low and thus lead to unbalanced training and the convergence of the loss function would be harder. However, it should be noted that adding a variability of ±10km on the depth allows us to actually go down to 70km. We clarified it in the revised manuscript, please refer to answer to question #12.

8. The ML algorithm uses in the end specific features to detect SSEs. Can the authors retrieve what are the important features to detect SSEs?

We agree with the reviewer that this would be very interesting to better understand how it works in detail, yet the assessment of the most important features is a complex task that goes beyond the scope of the manuscript, which needs to be intended as a "proof of concept". Going into the features computed by this specific network is not trivial because of the pooling and the transformer module, and also relevance propagation methods (to assess the most important features and relevant parts of the data, that the network looks at) are not easy to adapt. We keep this suggestion for a future development.

While it's important to answer those comments, I do not think that the modifications they require will change drastically the main results of this study. In this respect, I suggest a major revision to allow the authors some time to provide the additional content. You'll find below additional comments. Some of them will detail the main comments above.

Additional comments:

1. (P2) l.2: Ref for the duration of SSEs (e.g. Behr and Burgmann 2021).

Thank you for the comment, we added the reference (Behr and Burgmann, 2021) to the revised manuscript.

2. (P2) l.23: Tremors and LFE are not explained before. A sentence mentioning that they are micro-seismicity which are often occurring during SSEs in certain subduction zones is sufficient.

Thank you: we added this sentence to the revised manuscript (lines 24-26): "Tremors and LFEs are weakly emergent micro-seismicity, or micro seismicity with low frequency content often accompanying SSEs in certain subduction zones (Obara, 2002; Rogers and Dragert, 2003)".

3. (P2) l.24: What do you mean by "in-situ" geophysical monitoring? It's not clear.

We mean "in place": geophysical deployments on top of the targeted fault zone.

4. (P2) l.31: "tackle" instead of "tackled"?

Yes, thank you. We corrected the typo in the revised version of the manuscript.

5. (P2 and P3) l.39 and l.48: “end-to-end”? What do you mean by "end-to-end"?

An “end-to-end” model is defined as a model that does not require intermediate steps (e.g., manual intervention), but can be thought as a pipeline that takes some inputs (the first “end”) and does all the processing all the way to the output (the other “end”). We clarified it in the revised manuscript: “In order to develop an end-to-end model (which does not require any manual intermediate step)”.

6. (P3) l.52: I suggest removing the word “systematically”. There is always a possibility that the method do not detect small SSEs.

The reviewer is right, we removed this word in the revised version of the manuscript.

7. (P3) l.66: It’s not directly clear what is done here. It is only after reading the method section that the reader begins to understand the procedure and make sense of this paragraph. Maybe it’s the vocabulary that loses the reader. For example, the word “sets” in “sets of geodetic time series” is too vague and other words like “samples“, “segments”, or “portions” might better to illustrate de procedure. Similarly I did not understand what “single sample” in l.67 really referred to. By changing the word “set” to “sample” and changing the paragraph accordingly, it might clarify the procedure.

Thank you for the remark. The reviewer is right, the technical jargon can confuse the reader. In the revised manuscript, we changed “set of signals” into “group of signals” and “single sample” to “single input unit”, hoping that this would clarify the overall procedure.

8. (P3) l.66: Why take a 60-day window? If there are 2 SSEs in 60 days, can the method detect the 2 events?

We chose a 60-day window to be able to correctly detect a slow slip event with duration ranging from 10 to 30 days (for the longest one, the method can rely on the pre- and post-SSE noise to “see” its full evolution in the GNSS time series). Of course, training with one event per window is still a limitation of the model. However, when applying SSEdetector with a sliding-window mode, two events can be retrieved, as shown in Supplementary Figure 21. Please also refer to the answer to question #3 (major).

9. (P3) l.72: “must”? Is the ratio between time series with and without synthetic SSEs important? Why must it be half?

Thank you for your comment. What is important is not the ratio, but the fact that there are two parts, one of which must be made of only noise, and the other one must contain also some signal. In the text, we say: "On the other hand, while half of the samples (*negative* samples) will only consist of background noise, the other half must also include an SSE signal". We revised it at lines 109-112: "while a subset of the samples (*negative* samples) will only consist of background noise, another subset must also include an SSE signal. Here, we use a training database containing half positive and half negative samples."

10. (P3) l.76: "ultra-realistic"? It's a bit over the top.

Yes, we agree with the reviewer that this may seem "over the top", and we decided to remove it by keeping "realistic noise" in the revised manuscript.

11. (P3) l.80: Specify each time which sub-section in the "Method" section?

Thank you, we modified the text by referring to the specific sub-section at each time.

12. (P4) l.89: The authors use the fault geometry Slab-2. If you change the fault geometry, does it change something to the detection? I am mostly thinking about the southern region. The authors stop their fault at 60 km depth, while tremors go deeper than that (Fig1.b). It's a bit surprising since, in Cascadia, SSEs are occurring within the tremor band. Maybe it's a problem of the fault geometry at this location which is more difficult to assess. There are, I think, other studies that provide the geometry of the Cascadia subduction zone, and the authors can assess the uncertainty on the geometry.

Thank you for the comment. We have analyzed the problem and the issues that could arise from a specific slab model. This is the reason why we add a variability of ±10km to the depth (please see answer to question #7). Adding this variability can allow us to better generalize over different slab models and, most importantly, on real GNSS data. We added the following in the revised manuscript at lines 133-135: "Such a variability on the depth allows us to better generalize over different slab models and, most importantly, on their fitting to real GNSS data.".

13. (P4) l.91: What's the depth range? Are potential SSEs near the trench, between 0 and 20 km depth, taken into account? Reading the method section, it seems so, but this small precision needs to be clear.

Thank you. We modified the sentence "Their depths follow the slab geometry and are taken down to 60 km" to: "Their depths follow the slab geometry and are taken from 0 to 60 km, with further variability of ± 10 km".

14. (P4) l.92: I suggest removing "realistic".

We agree with the reviewer that modeling a realistic stress drop for SSEs is still difficult. Therefore, we removed the word from the revised version of the manuscript, as "We further assign each event a stress drop modeled from published scaling laws […]".

15. (P4) l.96: Why not more than 30 days. There are SSEs that last longer than 30 days? I acknowledge though that the SSEs in Cascadia propagate and a point on the fault

might not experience a SSE for more than 30 days. Same question for 10 days. That might hinder the detection of smaller events that will have shorter durations.

Please refer to answer to question #3 (major).

16. (P4) l.97: Is the amplitude of the synthetic SSEs slip distribution homogeneous along-strike? Does it matter? There is something unclear for me: the ML algorithm has no information about the precise location of the GNSS stations but it still learned to recognize a SSE if certain stations moved together due to a common SSE source defined by the synthetics. There's then an information of the connectivity of the GNSS station and their sensitivity to SSEs. Am I correct?

Yes, the slip distribution is homogeneous along strike (cf. Okada's (1985) dislocation model). It does not matter at first order since the model learns how to relate some transients in the data to slow slip events. We agree that having more realistic sources might improve the performance but we think that at first order, for the detection problem is fine. About the second question, yes, thanks to the spatial pooling, the method relies on observations coming from multiple stations (multi-station approach), i.e., it is not "triggered" by a single-station recording, but the slow slip signal needs to be coherent along more than one station, which we think is one of the added values of our work. Please also refer to answer to question #3 (major).

17. (P5) l.124: I suggest changing the sentence "This is done by assigning a weight to the data, with those weights being learnt from the data itself" to "This is done by assigning a weight to the data, with those weights being learnt from the data itself and the a priori knowledge of the labels, i.e. whether there is a SSE or not.". Is that correct? Otherwise it feels a bit like magic.

Thank you for the question. We agree with the reviewer that it is not magic at all, just mathematics. The weight assigned by the transformer (self-attention) is computed by building a similarity matrix and, for each day sample, finding the most relevant samples in the data that have strong similarity. This is then "optimized" during training. We think that going into the details of how a Transformer works would be a bit too complex. Therefore, we modified the text of the paper by incorporating your suggestion. Please find the revised text at lines 167-170: "This is done by assigning a weight to the data, with those weights being learnt by finding significant connections between data samples and by relying on *a priori* knowledge of the labels, *i.e.*, whether there is a SSE or not".

18. (P5) l.125-127: "identify the precise timing of aseismic deformation transients"? But are they not always position at the center of samples?

The reviewer is right: the synthetic SSEs are positioned in the middle of the synthetic time series. Yet, it is important to notice that a machine learning method does not "know" *a priori* where this information is. This is why we use a Transformer neural network: to find slow slip hidden in the noise. Even if we train it with SSEs always at the center, the Transformer learns how to effectively separate them from the noise, which translates into an effective retrieval of slow slip in real data also in more complex scenarios (more than one event, event migrations, etc.).

19. (P6) Fig.2: There are events with magnitude below 6 while it is said that the range of magnitude tested is in between 6 and 7. Why so?

Thank you for the comment: we were imprecise while describing that. We use magnitudes from 6 to 7 to have a balanced training (most events smaller than M_w 6 likely look like noise: in this case the loss function does not converge to a satisfactory minimum). However, we test the performance of the method from magnitude 5.5 in order to study its sensitivity. It should be noted that, since we generate random magnitudes and depths, some small events can still be retrieved. We added the following sentence in the revised manuscript at lines 190-192: "We generate events down to magnitude 5.5 to better constrain the detection limit at low magnitudes".

20. (P6) l.147: Refer to Fig2.a.

Thank you. We added the reference in the revised version.

21. (P6) l.153 and Fig2b: The magnitude threshold is interesting. I would show again the contour of the tremor band in Fig2b as in Fig1.b so that one can compare it with the threshold distribution. However, considering this slab geometry, the tremors seem to go deeper than 60 km depth. Why not take deeper synthetic events (e.g. down to 80 km depth)?

Thank you for the question. For a complete answer, please refer to the answer to (major) question #7 and (minor) question #12.

22. (P7) l.170: The detector is smoothed in time and the size of the window potentially controls how much it is smoothed. Does changing the window to 20/80 days change this smoothing? Additionally, it is a detector over the whole Cascadia subduction zone. What if there are 2 SSEs (e.g. one in the north and one in the south) happening at the same time?

In the answer to question 4, we proved that the choice of the window as well as other parameters does not produce any significant temporal smoothing. Yet, we thank the reviewer for the comment about two SSEs in the same window. In this case (e.g., one event in the north and one in the south), SSEdetector is able to identify that slow slip is found but, as we wrote in the discussion and conclusion, one current limitation is that the model does not localize the slip: for this reason, we cannot know, *a priori*, where the rupture is situated and, unfortunately, how many of them there are. We show this in Supplementary Figure 22. For more details, please refer to answer to question #3 (major).

23. (P7) l.182-184 and Fig.2 in supplement: The actual location of the events detected by Michel et al. (2019) are further inland, within the band of tremor activity, than what is shown in Fig.2 in the supplement. The magnitude detection threshold should then a bit higher than 6.5.

Thank you for the remark. We corrected the figure accordingly.

24. (P9) Fig.4.b: What's the black rectangle? It's not mentioned in the caption.

Thank you for the comment. We corrected the figure caption by adding: "The black rectangle represents an example of a large event found by Michel et al., that is split into three sub-events by SSEdetector.".

25. (P9) l.210: The authors can refer to Table 1.

Thank you. We modified the sentence "We present the duration distribution in Figure 4" into "We present the duration distribution in Figure 4 as well as a summary in Table 1".

26. (P9) l.215: The authors mention that they detect events shorter than 10 days but they did not train their detection method with synthetic events sorter than 10 days. Can the authors comment on this?

Thank you for your question. Please refer to answer to question #3 (major).

27. (P9) l.217-218 and Fig4: Note that Michel et al. (2019) provides minimum and maximum estimates of the duration. In any case, those estimates are biased by the some temporal smoothing that will tend to extend their duration. This comparison should be seen in light to this effect. The durations the authors provide might also be affected by a certain temporal smoothing which is not documented in the manuscript. The authors can take advantage of the synthetics to estimate the smoothing effect, which will also be linked with the detection probability threshold to define a SSE (i.e. 0.5 in this study). What is a good detection probability threshold considering the synthetics? How are their duration estimates compared to their true duration? The smoothing effect might be dependent of an events magnitude and/or on the size of the window. The authors can maybe produce a kernel that corrects the initial duration estimate.

Thank you for the comment. Please rely on the answer to the question #5 from reviewer #1 for a complete explanation. About the last question: we agree that this would absolutely be a good idea, but this correction would be very difficult to make since it is quite difficult to assess when and where the estimate of the duration would be biased by the SNR, the location of the SSE and other parameters. In our work, the duration estimate has to be taken just as a first order analysis and a proof of concept, without overinterpret it. A correct duration estimate is something related to the characterization of SSEs, which we take as a perspective work.

28. (P9) l.231 and Fig4: The authors comment on the points below the diagonal but not above. They detected events that have long durations (~70-80 days) but that are corresponding to short events in Michel et al. (2019). Can the authors comment on it? Is it due to the detection threshold or that SSEdetector cannot distinguish between 2 events if they occur at the same time?

Thank you for the question. First, rely on the explanation to question #22. This is due to the fact that the model cannot distinguish between two events occurring at the same time. Please refer to answer to question #3 (major). We clarified it in the revised manuscript at lines 306-309: "Also, a few points above the diagonal represent events that were split into sub-events by Michel et al., which our model tends to detect as a single event. This separation is also related to the choice of the threshold.".

29. (P11) l.247: I don't see 11 events in Fig3.b.

The reviewer is right: they are 7. We corrected the typo in the revised manuscript.

30. (P11) l.263-266: I am a bit concerned about the potential smoothing effects which is of relevance to those statements. Please see comment "(P9) l.217-218 and Fig4" for more details.

Please refer to the answer to question #28.

31. (P11) l.283 and Fig6 in supplement: Indicate in the caption which version of SSEdetector is the red and blue curve.

Thank you for the remark. We corrected the figure caption.

32. (P11) l.285: It's a bit surprising since there are more GNSS stations and thus potentially a better characterization of SSEs by the data. I guess there is a tradeoff between the number of station and the number of missing data.

Yes, the reviewer is right, there is a tradeoff between the number of GNSS stations and the number of missing data. We clarified this in the revised manuscript: "the detection power slightly decreases, probably due to a larger number of missing data, suggesting that there is a trade-off between the number of GNSS stations and the number of missing data.".

33. (P11) l.287-298: The results for the southern region, in the supplement, would be also nice.

Thank you for the comment. We added the results for the southern part in supplement.

34. (P11) l.349-352: Doubts remain considering potential smoothing effects. See comment "(P9) l.217-218 and Fig4" for more details.

Please refer to the answer to question #28.

Method: I am not an expert in Machine learning methods. Consequently I do not really understand all the points made regarding the method. My questions may arise as a non-expert in this field.

35. (P13) l.366-369: "…applying random transformations…" I do not understand the significance of this sentence. Is it necessary? Or does it imply an underlying bias? It's not clear to me.

Thank you for the comment. We clarified the sentence from "synthetic data is performed by applying random transformations" to "synthetic data is performed by applying a methodology based on randomization (SSEgenerator)".

36. (P14) l.374: What is the longest data gap sequence in a 60-day window?

By choosing 135 stations, the longest data gap sequence in a 60-day window can also be 60 days. By taking the "worst" station among the best 135 (station SQMS, located in the Vancouver Island), the longest data gap in the full time series is 124 days, which means that some synthetic samples will just not have any data from that station.

37. (P14) l.374-377: State why do you train also on 352 stations. I assume it's for a comparison that will be in the discussion. It was not totally clear at first read.

Thank you for the comment. Yes, it is for comparing with Michel et al (2019). We changed the sentence in the revised manuscript from "we also train and test SSEdetector on 352 stations (the same number used in the study by Michel et al. (2019))" to "we also train and test SSEdetector on 352 stations (the same number used in the study by Michel et al. (2019)) to compare with a setting equivalent to the one from Michel et al. (2019)".

38. (P14) l.400: "experimentally"?

Thank you for noticing that. We changed that in the revised manuscript from "The number of AAFT iterations has been experimentally set to 5" to "The number of AAFT iterations has been set to 5 based on preliminary tests".

39. (P14) l.403: Why 70% of the data? Why is it necessary to keep 30% in place? It's not clear. How do the results change if you change this ratio to a 100%.

Thank you for the comment. We chose to have the 30% of the data not affected by data gaps to help the model while training to fully recognize what a synthetic SSE look like. This is not mandatory but we found that it helps the training. We chose to have the 30% based on preliminary analyses on the convergence of the loss function during training.

40. (P14) l.407: I would change "…in order not to…" to "…to avoid…". It appears a few times in the manuscript (e.g. l.415).

Thank you, we rephrased that in the revised manuscript.

41. (P14) l.410: "…SSEdetector…" instead of "…SSedetector…"?

Thank you for spotting this typo, we corrected it in the revised manuscript.

42. (P14) l.414-420: Not clear. Up to this line, I understand the authors have built surrogate data of a size similar to X and that 70% of the data gaps were shuffled from their initial positions. Are the authors cutting X_R into contiguous sections of size W_L? I am not sure since 60,000 synthetic samples are generated. Or does it mean that there will be duplicates in the 60,000 samples? If X_R is not cut into sections, I do not follow the logic. I suggest the authors explain at the beginning of the section a broad guideline of what they intend to do to help the reader follow.

Thank you for the comment. Yes, we cut X_R into contiguous (**non-overlapping**) windows of size W_L. Since we take contiguous windows, the final data set will not contain any overlapping window nor any duplicate. In fact, we cut a certain number of non-overlapping windows for each "noise generation step", repeating it several time to reach 60,000 windows.

43. (P15) l.415: Since the data gaps where already shuffled X_R, why apply again a circular shift? After reading the whole manuscript, I think I understood. But when I read this sentence it was not obvious at all.

The reviewer is right, we were not precise in the description. In fact, the data gaps are shuffled in the sense of the stations and the circular shift is made in time. We modify the text accordingly: "We shuffle the data gaps before imprinting them to the data, assigning a station a data-gap sequence belonging to another station, such that […]" and "the data is circularly shifted, in time, by […]".

44. (P15) l.423: I guess the authors are selecting randomly the barycenter of the SSE when they mention its location?

Yes. We corrected “random locations” to “random locations (fault centroids)”.

45. (P15) l.432: Why not smaller than 6? Besides, Fig 2.a has events lower than 6.

Please refer to the answer to question #19.

46. (P15) l.433-434: Instead of “fault geometry” or “fault radius” I suggest SSEs geometry and radius.

Thank you, we followed the advice and we corrected it in the revised manuscript.

47. (P15) l.433-440: There might be a typo in the equation for the average slip as it does not depend on a characteristic length. Stress drop is often approximated using the following equation (Kanamori and Brodsky, 2004; Noda et al., 2013):
    $(\Delta\sigma) = C\, M\_0/\rho\hat{}3$,
    where M_0 is the moment released by an earthquake, ρ is a characteristic length of the rupture and is taken as ρ=A^(1/2), A corresponds to the area of the rupture, and C is a geometrical parameter which depends on the shape of the rupture. According to Noda et al. (2013), C is equal to 2.44 for a circular crack, and C= 2.53, 3.02 and 5.21 for rectangular ruptures with aspect ratio of 1, 4 and 16, respectively. In parallel, M_0=μAu, where μ is the shear modulus and u the average slip of the event. Thus
    Δσ =(CμAu)/A^(3/2) =(Cμu)/A^(1/2) ,
    and
    u = (Δσ A^(1/2))/Cμ.
    The average slip depends on the area contrarily to the equation written in the manuscript, and assuming an aspect ratio with W=L/2, it follows
    u = (Δσ(WL)^(1/2))/Cμ=(ΔσL)/(√2 Cμ).
    The authors could calculate C for an aspect ratio of 2, or take C between 2.53 and 3.02 (i.e. aspect ratio of 1 and 4). In any case, I do not think the value of C will have a strong impact on results, but the equation of slip should be corrected.

Thank you for the remark. We corrected the equation (cf. answer to question #2). For the remaining questions, please refer to answer to question #3.

48. (P15) l.439: The aspect ratio of 2 seems small for SSEs in Cascadia. Michel et al., 2019 implies that M_w>6 SSEs in Cascadia have aspect ratios between ~3 and 13. While there are biases when inverting for slip on the fault, I do not think they will change to first order the fact that SSEs in Cascadia have large aspect ratios. How does the

aspect ratio of 2 impact the results? Are actual detected SSEs cut into multiple events? What does it change if the authors choose an aspect ratio of 4?

Please refer to answer to question #3.

49. (P15) l.444: Why a lognormal? Using the values provided by the authors, we obtain a stress-drop pdf as showed in Figure R1. Is that the PDF chosen?

Yes, this is the chosen PDF. We use a lognormal law as it has been shown to be a good probability density function for earthquake modeling. Please find below a few references:

Allmann B. P., and P. M. Shearer (2007), Spatial and temporal stress drop variations in small earthquakes near Parkfield, California, J. Geophys. Res., 112, B04305

Allmann B. P., and P. M. Shearer (2009), Global variations of stress drop for moderate to large earthquakes, J. Geophys. Res., 114, B01310

Balltay A., S. Ide, G. Prieto, and G. Beroza (2011), Variability in earthquake stress drop and apparent stress, Geophys. Res. Lett., 38, L06303

Oth A., D. Bindi, S. Parolai, and D. Di Giacomo (2010), Earthquake scaling characteristics and the scale-(in)dependence of seismic energy-to-moment ratio: Insights from KiK-net data in Japan, Geophys. Res. Lett., 37, L19304

Shearer P. M., G. A. Prieto, and E. Hauksson (2006), Comprehensive analysis of earthquake source spectra in southern California, J. Geophys. Res., 111, B06303

50. Figure R1: Stress drop PDF using c_V=10 and (Δσ) =5 10^4 Pa. => Please see PDF

Please see previous answer.

51. (P16) l.455-462: The term “steady state value of the signal (i.e. D) ” is a bit confusing since the signal is also steady state after the SSE (the position values are just close to zero).

Thank you for the remark. We changed the “steady-state” term into “post-SSE” in the revised manuscript.

52. (P16) l.459: t_max should be linked to a displacement of Dγ, and t_min to D-Dγ (the opposite of what is said in the manuscript). Note that I have written D-Dγ instead D-γ since I interpreted γ as a percentage of the signal amplitude. This implies that $\beta=2/T \ln((1-\gamma)/\gamma)$. If I was plotting ds(t) taking the equation of β from the authors and trying to generate a SSE with T=30 days and D=10, the SSE will not have a duration of 30 days. It will be instead around 18 days. By doing the modifications as proposed, the authors will correct this problem. Besides, I suggest to the authors to change/remove ds(D-Dγ) and ds(Dγ) at l.460-461 as ds is written as a function of time in equation (5). It’s at first sight a bit confusing even though ds is also a function of D. Also, there is an underlying assumption that the loading of the fault is disregarded since the “steady state” slope is 0.

We thank the reviewer for the comment. Please refer to the answer to comment #2 for the detailed response.

53. (P18) l.545 and Figure S15: The actual signification of h_L is confusing when looking at the supplementary Figure 15. The topographic prominence is defined by the difference between the height of the peak and the height of its highest base: p=h1-B2. Thus h_L=h1-0.7 p is a height slightly above B2. In the supplementary Figure 15, h_L is represented as a slightly shorter version of p. Additionally, in this figure, there are two steps with the number "3" and the text/figure could be bigger. Finally, some examples using real data, showing the number of tremors per day and the estimated final width, would be nice.

Thank you for the remark. We corrected the figure accordingly.

54. (P19) l.568: What is p''(t) and n''(t) ?
The authors should give in the supplement the timing and duration of each SSE they detected. They should also provide at some point (e.g. external files) the probability of SSE detection (Fig3.a) and other data necessary for plotting Fig4.b and 5.d.

Thank you for the comment. It was a typo from a previous version of the manuscript. We removed it. About the data: yes, we will upload our catalogue on GitLab, where the reader can also find all the data and codes to reproduce all the results and figures.

29th Sep 23

Dear Mr Costantino,

Please allow us to sincerely apologise for the long delay in sending a decision on your manuscript titled "Slow slip detection with deep learning in multi-station raw geodetic time series validated against tremors in Cascadia". It has now been seen by our reviewers, whose comments appear below. In light of their advice we are delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Earth & Environment under the open access CC BY license (Creative Commons Attribution v4.0 International License).

We therefore invite you to revise your paper one last time to address the remaining concerns of Reviewer #4. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table".

*****Please take care to match our formatting and policy requirements. We will check revised manuscript and return manuscripts that do not comply. Such requests will lead to delays. *****

Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed at the end of the Editorial Requests Table; the list of required files is also available at https://www.nature.com/documents/commsj-file-checklist.pdf .

OPEN ACCESS:

Communications Earth & Environment is a fully open access journal. Articles are made freely accessible on publication under a CC BY license (Creative Commons Attribution 4.0 International License). This license allows maximum dissemination and re-use of open access materials and is preferred by many research funding bodies.

For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit https://www.nature.com/commsenv/article-processing-charges

At acceptance, you will be provided with instructions for completing this CC BY license on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

Please use the following link to submit the above items:
[link redacted]
** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Teng Wang, PhD
Editorial Board Member
Communications Earth & Environment
0000-0003-3729-0139

Joe Aslin
Senior Editor,
Communications Earth & Environment
https://www.nature.com/commsenv/
Twitter: @CommsEarth

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

My previous comments have been reasonably addressed.

Reviewer #4 (Remarks to the Author):

The authors present a study where they construct a synthetic GNSS dataset containing slow slip events signature (with Cascadia subduction zone reference) to train an AI to detect SSEs in real-time series in Cascadia.
They compare their results with previously published SSEs catalog constructs using GNSS time series and ICA method and with a tremor catalog.
I found the manuscript well-written and clear. The authors answered well to the previous comments and made the necessary modifications in the manuscript. If I am unfamiliar with deep learning, I have learned from reviewing this paper.

I recommend the paper for publication after a few minor comments.

1. I guess the study is made by using daily GNSS positioning. However, you do not mention it in your manuscript; it is worth saying it clearly. I wonder about it since I found that being able to detect SSEs within a 2-day duration is questionable by using daily GNSS data. What are the incertitudes about short-term events?

2. You mentioned that several SSE catalogs have been made in this region, but you compare your results only with one record. Why?

3. I found the comparison between your catalog and Sylvain Michel's catalog interesting. It highlights the limits of ICA. However, to determine SSE duration, you must establish a threshold. It would be great to compare the cumulative amplitude of each event (for split events from Michel, you might sum them?). I expect you to find a difference because of this threshold, and I wonder how representative this comparison is for longer SSEs. I am not hoping that Michel finds short SSEs due to the smoothing they apply.
You could also add a few sentences in your discussion about that.

4. The overall strategy you added helped a lot. However, it could be more precise. For example, how many stations do you use? 1? a network? A sub-network? By samples, do you mean time? If yes, what is a sample? An hour? a day?
L91: you did not refer Michel et al. to a citation, Michel et al. (2019) [X]. There are other lines with the same issue.
L92: "different signal processing technique." you could be more precise: temporal ICA (using X components).

L99: with conventional methods: which ones? Are there other catalogs than Michel's one?

L109 - 115: Do your SSEs generated migrate over space?
After Reading the whole paper, I found the answer; however, it would help to add a sentence here to help the reader.

For each event you create, do you constrain the slip direction?

Minor revision

In the paper, you sometimes write geodetic time series and GNSS time series. I'd suggest using only the GNSS time series (as you do not use any other dataset) to clarify it.

Figure 1. Do you have the resolution to recover SSEs within Mw 7 near the trench? Maybe you could add some transparency in regions where the answer is poor. Did you test how your model is sensitive about events near the trench?

L272-275: add a figure reference.
L280" "many more": how many? Can you be more precise?

Dear Editors,

Thank you for the comments received on our article. In this version, we respond to the general comments raised by Reviewer #4. Their questions are colored in black and our answers are reported in blue. Each modification to the text of the manuscript is shown in green.

We hope that the current version will convince you, so that the paper can be accepted for publication.

Sincerely,
Giuseppe Costantino on behalf of the co-authors

Reviewer #4 (Remarks to the Author):

The authors present a study where they construct a synthetic GNSS dataset containing slow slip events signature (with Cascadia subduction zone reference) to train an AI to detect SSEs in real-time series in Cascadia.
They compare their results with previously published SSEs catalog constructs using GNSS time series and ICA method and with a tremor catalog.
I found the manuscript well-written and clear. The authors answered well to the previous comments and made the necessary modifications in the manuscript. If I am unfamiliar with deep learning, I have learned from reviewing this paper.

I recommend the paper for publication after a few minor comments.

1. I guess the study is made by using daily GNSS positioning. However, you do not mention it in your manuscript; it is worth saying it clearly. I wonder about it since I found that being able to detect SSEs within a 2-day duration is questionable by using daily GNSS data. What are the incertitudes about short-term events?

The reviewer is right, we did not specify that we are using daily GNSS positioning. We added this in the introduction and in the Results: "daily GNSS position time series".
About the uncertainty on short-term events: this is reflected by the probability value, that can be also interpreted as the confidence of the model. For some short-term events this value is often around 0.5, also suggesting that these results have to be taken with caution, since these events hardly emerge from the noise.

2. You mentioned that several SSE catalogs have been made in this region, but you compare your results only with one record. Why?

According to (Ide and Beroza, *PNAS*, 2023), the SSE catalogues from (Schmidt and Gao, *JGR*, 2008) and (Michel et al., *Nature*, 2019) are available for the Cascadia subduction zone. The catalogue from Schmidt and Gao spans the period 1998 – 2008, the catalogue of Michel et al. covers 2007-2017. Since we analyze data from 2007 to 2022, we thus compare our method with Michel et al., their catalogue being, to our knowledge, the most complete and appropriate for our date choice.

3. I found the comparison between your catalog and Sylvain Michel's catalog interesting. It highlights the limits of ICA. However, to determine SSE duration, you must establish a threshold. It would be great to compare the cumulative amplitude of each event (for split events from Michel, you might sum them?). I expect you to find a difference because of this threshold, and I wonder how representative this comparison is for longer SSEs. I am not hoping that Michel finds short SSEs due to the smoothing they apply. You could also add a few sentences in your discussion about that.

Michel et al. choose this threshold based on the amount of slip. In our case, this is not directly comparable, since we use a threshold on the probability of detection. However, the latter cannot be used as a proxy for the slip, which complicates a bit the comparison. In Figure 4b, we perform a first-order comparison of our durations and those of Michel et al. The events from Michel et al. are, in some cases, aggregated as longer events by our method and, conversely, longer events by Michel et al. are rather split by our method in sub-events. Nevertheless, the duration estimation is always associated with a threshold that must be chosen, which is not straightforward and is always influenced by a-priori choices (Ide and Beroza, *PNAS*, 2023). As the Reviewer is pointing out, longer events are affected a lot by the choice of the threshold, and an even comparison becomes harder.

4. The overall strategy you added helped a lot. However, it could be more precise. For example, how many stations do you use? 1? a network? A sub-network? By samples, do you mean time? If yes, what is a sample? An hour? a day?

Thank you for your feedback. We added the missing information to the section "Overall strategy". We replaced "SSEdetector, which detects slow slip events in GNSS time series" with "SSEdetector, which detects slow slip events in **135** GNSS time series". We rewrote "We first create 60,000 samples of synthetic GNSS position time series" as "We first create 60,000 **(60-day)** samples of synthetic GNSS position time series".

5. L91: you did not refer Michel et al. to a citation, Michel et al. (2019) [X]. There are other lines with the same issue.

Thank you, we corrected these typos.

6. L92: "different signal processing technique." you could be more precise: temporal ICA (using X components).

We changed "different signal processing technique" into "different (independent-component-analysis-based) signal processing technique" in the revised manuscript.

7. L99: with conventional methods: which ones? Are there other catalogs than Michel's one?

Thank you for the remark. The sentence is not clear. We corrected it from "a preliminary catalog of SSEs has recently been proposed during the period 2007-2017 by Michel et al. with conventional methods" to "[…] catalog of SSEs has recently been proposed during the period 2007-2017 by Michel et al. based on time series decomposition via independent component analysis" in the revised manuscript. For the second question, please refer to question #2.

8. L109 - 115: Do your SSEs generated migrate over space? After Reading the whole paper, I found the answer; however, it would help to add a sentence here to help the reader.

We added a sentence "The modeled signals do not propagate over space", at lines 115-116.

9. For each event you create, do you constrain the slip direction?

Yes, we constrain the synthetic faults to have a thrust focal mechanism. The Reviewer can find this information in the text at lines 130-131 "[…] the focal mechanism of the synthetic ruptures approximates a thrust […]".

Minor revision

10. In the paper, you sometimes write geodetic time series and GNSS time series. I'd suggest using only the GNSS time series (as you do not use any other dataset) to clarify it.

Thank you for the suggestion. We modified accordingly any occurrences where "GNSS" can better replace "geodetic".

11. Figure 1. Do you have the resolution to recover SSEs within Mw 7 near the trench? Maybe you could add some transparency in regions where the answer is poor. Did you test how your model is sensitive about events near the trench?

The reviewer perhaps refers to Figure 2. In that case, yes, a Mw 7 near the trench can be detected along the whole subduction (more precisely in southern Cascadia).

12. L272-275: add a figure reference.

Thank you for your comment. We added a figure reference at lines 274-275: "The shape of the probability curve gives insights into how SSEdetector reveals slow slip events from raw GNSS time series **(see Figure 3)**."

13. L280" "many more": how many? Can you be more precise?

We detect 43 additional events. We corrected the text in the revised manuscript.