To all reviewers, thank you for your considered comments and feedback.

Reviewer 1. Regarding *originality, impact* and *novelty*, although similar constructions could be applied to other GP approximation schemes, the sparse spectrum GP in particular is capable of closed-form (input) uncertainty propagation with regard to the latent warping. No sampling is needed, unlike say DGPs. This consequently admits the results we have given in the paper (e.g. eq. 11 & 12). Furthermore we believe the simplicity of the warping is a major feature from the practitioner's perspective.

About the presentation aspects, on *terminology*, the presented formulation is agnostic to the degree of differentiability. From the kernel perspective, as seen in the experiments, we have used finitely differentiable kernels like the Matern 3/2. It would certainly be interesting as future work an analysis of kernels, like the pure exponential, however this was not within the scope of this work. Lastly, regarding *structure*, we will move the related work to an earlier section.

Reviewer 2. Thank you for the kind comments and additional feedback!

Reviewer 3. *On injectivity*, we understand that an injective warping would allow us to map points in $\mathcal{Q}$ back into the original $\mathcal{X}$. It is also noted that the deep GP formulations based on dynamical systems may allow maintaining injectivity via monotonic constructions. Our stacked formulation of the SSWIM kernel is definitely not, in its current form, based on an injective warping. Injectivity (and even bijectivity) could be enforced as an additional constraint and undoubtedly deserves a dedicated work. An effect we can foresee is that the Gaussians given by each GP may collapse as the number of layers increases. However, we propose the stacked formulation not with the intention of applying a very deep transformation, but more towards shallow ones, as we analysed in Sec. 4.1.3. At this depth, the mentioned effects should be negligible.

An interesting sidenote is that one could argue injectivity is not necessarily ideal for learning latent mappings. It certainly is not a *necessary* condition for preventing collapse of uncertainty although such phenomena may be correlated. That is to say, by relaxing injectivity it is plausible for two different input values in a prior warping layer to warp the same input location in the next layer. This is in fact a potentially desirable property – it suggests compressiblity of the input domain – in that there might be an underlying non-monotonic/non-stationary covariance function at play. Such expressiveness would not be able to be directly captured by a purely injective mapping. We are happy to add this discussion to the revised version.

Regarding *specific priors*, this is an interesting discussion however not the focus of the paper. Our methodology is, generally speaking, "kernel prior agnostic" in the sense that the nonstationarity is accomplished through the affine transformation. Of course the latent function kernels play a role - one could indeed use extremely expressive kernels like the stationary Spectral Mixture. However to restrict the space of analysis to the effect of the warping construction we aimed to minimise kernel discovery. *ARD kernels, dimensionality.* We actually use stationary ARD kernels in the sense of lengthscale for the warping layers. They however will never be able to capture nonstationarity unless composed like in SSWIM for example. Trans-dimensional mappings are a whole new question we are currently exploring; indeed we restrict our mapping to retain the same dimensionality as input since the intuition of the affine mapping is strongest in this sense. For *Fig 1.* in the experiment for this figure we do not explicitly optimise the warping layer's marginal likelihood *fit w.r.t. pseudo training points* although doing so is entirely possible. *Dynamic input warping.* We did not compare against this, although we have referenced in the paper, since it is a different modelling paradigm. We have provided comprehensive experiments alongside the most structurally relevant methods (i.e. DGP, DKL). Such analysis would definitely be valuable for future work and we are happy to mention so in the paper.

Reviewer 4. With respect to *complexity,* we have provided a computation complexity summary in the paper (lines 211–219). Comparison methods differ in complexity, but the worst case is for a fully non-parametric GP at $O(N^3)$. For our *overfitting analysis*, we believe it was a valuable contribution to promote overfitting analysis as it is often not considered in related works (and even the GP literature in general, especially with regard to the typical marginal likelihood loss, where it is often assumed to prevent overfitting). Some papers like Ton et al [2] and Lázaro-Gredilla et al [7] have dedicated sections to such an analysis, however, it is not typical. We did not provide a comprehensive overfitting comparison of other methods because this deserves a separate work in and of itself. *Related work.* We have moved the related work discussion to the beginning. *Notation.* Thanks for the comments; we will consider a clearer notation for the revision.