We thank the reviewers for their considered reviews, and are pleased there is unanimous agreement on accepting the paper. In particular, reviewers found the paper provides significant contributions to both the modelling and training aspects of Neural Processes, is well written, and includes extensive and convincing empirical investigation. Since submission we have made improvements to the paper. We first address reviewer feedback, then describe these.

Reviewer feedback. We thank all reviewers for their helpful and constructive comments. We do not have space to address them all here, but will do so in the paper. We appreciate R1's suggestion to move Fig 1 in the supplement and Algs 3 and 4 to the main body, and will apply this to the revision. R1 mentions our image experiments lack multimodality, but note that we provide a demonstration of multimodality in Figs 3c & 3d. Nevertheless, we find in practice that the ANP tends to produce more diverse image samples: we will add a discussion of this point. R1 asks why Eq 4 is not a valid ELBO. The correct ELBO is: $\mathbb{E}_{q(\mathbf{z}|D)}[\log p(\mathbf{y}|\mathbf{x},\mathbf{z})] - \mathrm{KL}(p(\mathbf{z}|D_c)\|q(\mathbf{z}|D))$. However, as $p(\mathbf{z}|D_c)$ is intractable, this is *approximated* by $q(\mathbf{z}|D_c)$ (as described in Garnelo et al. [11] Eq. 9), hence this is no longer a lower bound in a single consistent model. We will expand on this in the paper.

R3 asks if the method can be applied on more challenging tasks. The extension to, e.g. few-shot image classification is interesting, but will require several modifications. As such, we leave this for future work. We agree with R4 that a comparison with ANP and ConvCNP on the real-world experiment in Sec 5.3 would improve the paper. We did not include it as our focus in this subsection was on translation equivariant models with coherent samples. It is unclear how to apply the ANP effectively here, since the test regions do not overlap with the train region, and are of different sizes. Further, the ConvCNP would not produce coherent samples, and cannot be applied to sampling in Fig 4 or to Thompson sampling in Fig 5. R4 points out that the improvement in changing objective from $\mathcal{L}_{\mathrm{NP}}$ to $\mathcal{L}_{\mathrm{ML}}$ is more significant for ConvNP than ANP. We believe this is because ConvNP has more latent variables than ANP (5000 vs 128), and, as alluded to on line 181, $\mathcal{L}_{\mathrm{NP}}$ is more detrimental for models with more latent variables, since the KL-term in Eq 5 is more of a 'distraction' from the max-likelihood target. We will expand on this in the paper.

Heteroskedastic noise. We have made a minor technical improvement which leads to significant performance gains: changing the Gaussian observation noise from homoskedastic (hom. noise) $\sigma_y^2(\mathbf{z})$, to heteroskedastic (het. noise) $\sigma_y^2(\mathbf{z},\mathbf{x})$ in Eq 2. This change follows the findings of Le et al. [20], who demonstrate that het. noise improves performance for several NP models, and is in line with Kim et al. [14]. We emphasise this is a design choice that does not affect conceptual aspects of the paper or model. Moreover, we find this improves performance for both ConvNP and ANP *and* is simpler to implement. Hence *we have rerun all experiments with het. noise and updated the paper*.

The results are: i) All models now perform better. An example with two kernels is provided in the table, comparing each model with het. vs. hom. noise. The trends in the table generalise almost without exception in our experiments. ii) The ordering of log-likelihood of ConvNP and ANP is unchanged in all experiments (including images): ConvNP ($\mathcal{L}_{\mathrm{ML}}$) still outperforms ANP. iii) As before, ANP fails catastrophically when spatially extrapolating. iv) ANP and ConvNP both show tight predictions around training data. v) Other conclusions in the paper are unchanged. To summarise, this change provides a strict improvement to the paper, as it simultaneously simplifies the implementation while leading to significant improvements in performance in all settings considered in the empirical section.

In addition, we discovered a plotting error in Fig 2 in the submitted version, which as a result, depicted lower noise variance for the ConvNP than should have been. We stress that this only affected *plotting* of the ConvNP in Fig 2, and none of the surrounding results or analyses. Moreover, with the change to het. noise, both the ConvNP and ANP now collapse their uncertainty around the data, and all of our results are replaced with improved versions. An example is depicted in the figure below, where the top row details the corrected plots with hom. noise, and the bottom row shows the models' predictives with the new het. noise on the Matérn–$\frac{5}{2}$ kernel. These plots should be compared with the 3rd row of Fig 2 in the submission, which contains the plotting error for the ConvNP.

	ConvNP (hom.)	ConvNP (het.)	ANP $\mathcal{L}_{\mathrm{ML}}$ (hom.)	ANP $\mathcal{L}_{\mathrm{ML}}$ (het.)	ANP $\mathcal{L}_{\mathrm{NP}}$ (hom.)	ANP $\mathcal{L}_{\mathrm{NP}}$ (het.)
Matérn-$\frac{5}{2}$	$-0.80 \pm 7\text{E}{-3}$	$\mathbf{-0.58} \pm 0.01$	-0.78 ± 0.01	$\mathbf{-0.73} \pm 0.01$	$\mathbf{-0.95} \pm 8\text{E}{-3}$	$\mathbf{-0.96} \pm 0.01$
Sawtooth	1.22 ± 0.01	$\mathbf{2.30} \pm 0.01$	$-0.03 \pm 3\text{E}{-3}$	$\mathbf{0.09} \pm 3\text{E}{-3}$	$0.02 \pm 4\text{E}{-3}$	$\mathbf{0.20} \pm 9\text{E}{-3}$