We thank the reviewers for their positive feedback and suggestions. Below we address the major points raised. We will also be sure to incorporate the smaller suggested edits in the final revision.

Reviewer 1

Compared with DEQ, the fixed points of MOE are much more stable and unique, which needs experiments.
Thank you for pointing this out. We will definitely highlight this comparison more in the final version. The original DEQ paper (Bai et al. 2019) uses several training tricks to obtain stable fixed-point convergence, such as gradient clipping, training with subsequences, constraining the weight initialization, and warm start with a pretrained shallow network. There are also anecdotal accounts of difficulty obtaining convergence of DEQs, e.g. Linsley et al. 2020.

In addition, in response to this point we have some additional experiments which we will expand upon further in the final paper. Specifically, we attempted to train a "standard" DEQ with our multi-tier convolutional architecture using Broyden's method and W being unconstrained by the monotone conditions. Using this standard DEQ setup the models would become unstable and diverge after 1-2 epochs, even when using a very large (300) maximum number of Broyden iterations. We will include these results in the paper.

In Theorem 1, the function f should be convex closed proper (CCP).
Great point, we mention a bit of discussion around the CCP requirement in Appendix A but will be sure to add it to the theorem statement and discuss in the main paper. Note that all the activations we discuss (e.g. ReLU, tanh, sigmoid, and softplus) can in fact be represented (or approximated) as prox operators of CCP functions.

Reviewer 2

NeuralODE and DEQ extend to series data, but it is not clear how to extend MON to such data.
This is an excellent point. It *is* indeed possible to apply MONs to time-series data by using 1D causal convolutions. There are some subtleties involved in extending MONs to this setting (such as ensuring that influence remains causal when using FFT-based convolutions, which are naively circular), but these can be overcome by careful use of zero padding and other features. We will absolutely highlight this in the updated paper and include some simple experiments.

In the experimental section, it would be nice to see the results of NODE and ANODE for bigger model size along with the results of MON.
Absolutely. We ran these experiments over the rebuttal period. Although we hope to run further experiments to verify this behavior, our chief finding so far (which is consistent with discussions we have had about Neural ODEs with others), is that larger ODE models actually begin to *diverge* after a certain point in training. For example, the best performance we could obtain using NODE or ANODE models with ~1M parameters on CIFAR-10 is 72%, which still vastly underperforms MON.

Could the authors comment on how their approach compares to the "Implicit Deep Learning" paper?
Great point, and we will add further discussion about this to the paper. The short answer is that the conditions they require more directly relate to stability of the Piccard (simple forward) iteration, and are thus not directly comparable (sometimes the method will be stable while not monotone, and for some monotone operators the naive Picard iteration is not stable, even though other operator splitting methods will be). We will absolutely discuss these points more fully in the paper.

Is there any intuition on whether such architectures are suitable for regression tasks?
Yes, since MONs are fundamentally about constructing a fixed point as the hidden representation, we can use a last layer and loss function for regression tasks, just as with normal networks.

Reviewer 4

There are still some comparisons that are missing compared to the reported results in NODE and ANODE.
Hopefully our experiments with larger ODE models (described above) help to address this point. Please let us know if there are additional comparisons we should include in the final version.

It would be great if authors could report the standard error for their results similar to NODE and ANODE papers.
Thanks for pointing this out. We have added error bars to the convergence plots for the MON models, which are very narrow around the reported performance. Indeed, this is a major advantage of MON models, which appear relatively stable compared to the (A)NODE models.