Question on role of the presented application: We would like to clarify the aim of our work, as it explains the role of the chosen example application. The main motivation is to provide conceptual insight. Analytically unrolling recurrent dynamics into a (functional) Taylor series, where coefficients are given by Green's functions, is a versatile approach that may be used as a general purpose scheme. This expansion reveals how the non-linear interactions pick up higher order correlations in the input statistics, quantifying how non-linear networks provide a richer feature space than linear ones. The approach therefore elucidates which statistical features of the input data can be used by the network, thus opening a door to link and compare reservoir computing to feature-based approaches of classification.
The existence of an optimal input projection and readout vector furthermore allows one to first define and second study the performance of the recurrent reservoir itself. Thus, common methods for optimizing recurrent connectivity, including BPTT (suggested by reviewer 3), can be combined with our algorithm to study and improve the kernel properties of a reservoir network.
To illustrate the new concepts we chose one toy example that provides analytical insight and one real-world dataset to demonstrate practical applicability. We emphasize this view by the new title 'Unfolding recurrence by Green's functions for optimized reservoir computing' and agree with the reviewers that future work is required to systematically assess the performance on a broad set of problems; we are very grateful to reviewer 4 for proposing the Japanese vowel data set and the systematic approach described in Bagnall et al. 2017, that we plan to follow.

Question on input data: The input data does not need to fulfill any assumptions on stability, and in particular, as demonstrated with the ECG dataset, non-Gaussian stimuli are admissible. Section A.5 and A.6 in the appendix (Supplementary material) show how the soft margin is computed from the empirical moments of the stimuli, implemented in the provided code (to be published as a zenodo repository). The length of the input data is relevant for the computation of the Green's functions, which have to be computed only once before optimization and scale linearly ($G^{(1)}$) and quadratically ($G^{(2)}$) with the stimulus length. All summations over time indices can be performed prior to the actual optimization, which thus becomes independent of the length of the data. Centering of data is possible without knowledge of the labels and can also be avoided by incorporating a threshold in eq. (4) and following derivations. The scaling of inputs is arbitrary, as it can be absorbed by rescaling α. Both were performed here only for conceptual clarity, allowing identical network parameters for both datasets and to simplify the presentation of the mathematical details. The supplementary material furthermore contains a description of the full algorithm. The required level of detail to ensure reproducibility unfortunately prohibited the inclusion in the main text. We will point more clearly to the appendix and also include pseudocode in a revision.

Question on stability of dynamics, chaos, memory life time: While the Green's function of the linear system remains well-defined also in the linearly unstable regime (spectral radius of W exceeding unity; chaotic dynamics), the perturbative solution of the non-linear system built thereof (eqs. (9)-(12)) suffers from exponentially growing modes. We are currently working on a multi-timescale approach that propagates only for short temporal intervals and then recomputes the Green's functions; it appears to work reliably in the chaotic regime. Generally, the initial network state should not be chosen too large as the approximate solution of the Green's function requires operation in the near-linear regime; otherwise, it is irrelevant. The final network states do not converge to fixed points of the dynamics: If the network was evolved beyond $t > T$, the states would continue changing. In this regard, the presented work is drastically different from the view of computation by fixed-points and slow points (see e.g. review by Susillo et al. 2014), to be discussed in the revision.
The memory characteristic reviewer 3 was concerned with can be seen from the slowest decaying mode with time constant $\tau/(1 - \max \mathrm{Re}(\lambda_\alpha))$, where λ_α are the eigenvalues of W. For spectral radii close to unity, the time scale thus becomes very long (diverging); otherwise the network forgets exponentially in time. The statements by reviewer 2 on this topic remain true for optimized input projection. We pointed out the exponential forgetting, because in the presented applications the decision is made at the final time point. This means that there is room for improving performance in particular in the ECG dataset, either by using a reservoir closer to instability or by a different readout mechanism (e.g. integrated over time).

Question on non-linearity: For conceptual clarity we assumed a weak ($\propto \alpha \ll 1$) nonlinearity to enable a vanilla perturbative expansion. For stimuli with low linear separability, even small nonlinearities significantly increase classification performance (see artificial stimulus, Fig. 3b). This is, however, not the case for the chosen real-world application as the latter is already easy to separate linearly (giving rise to the same accuracies in Table 1, discussed in lines 235-36 and 241-44). Future work is left to assess the performance of weakly non-linear reservoirs on other real-world data. In addition, stronger, arbitrary non-linearities can be handled by the multi-timescale approach (see above), or by established methods: hard non-linearities (Heaviside; binary / spiking networks) allow the use of the Gram-Charlier expansion (Dahmen et al. 2016 PRX; Farkhooi et al. 2017 PRL), exploiting that intrinsically-generated network noise smooths out hard non-linearities, to be discussed in the revision.

Further comments: We will take care of the feedback by reviewers 1, 3 and 4 to improve the embedding into current literature and gratefully acknowledge also all minor points that are very helpful to revise the manuscript.