The state of the earth: estimating physical parameters from noisy and incomplete earth observation data

Arp, L.R.

7

GENERAL DISCUSSION AND CONCLUSION

We are now ready to look back at the research questions introduced in Chapter 1 and answer them. Over the last four chapters, we have introduced novel methods and analyses to improve parameter estimation from EO data. These contributions all tackled the problem from different angles, motivated by the particular challenges of parameter estimation using EO data (see Section 1.1):

- Challenge 1: Data inconsistency and spatio-temporal gaps (EO and ground truth)
- Challenge 2: Noise and ill-posedness on the inference problem

The answers to the research questions will contribute toward reducing the impact of these challenges.

7.1. ANSWERING RESEARCH QUESTIONS

We will answer the research questions, explained in Chapter 1, one by one.

7.1.1. RQ1: SPATIAL INTERPOLATION

The first research question we will answer is RQ1: *How can we effectively interpolate spatial data such that both local and global spatial properties are retained?* (Chapter 3)

For this research question we primarily focused on the spatio-temporal ground truth data (e.g., sensor network data) that is affected by Challenge 1. Existing methods were subject to many limitations, mainly regarding the tradeoff between modelling either local or global spatial relationships, and assumptions, including stationarity and isotropy. In Chapter 3, we proposed a novel spatial interpolation method, VPint (value propagation interpolation), capable of addressing these limitations and without assuming stationarity or isotropy. VPint incorporates a novel system-oriented perspective, as an alternative to the local- or global perspectives offered by existing methods. In this approach, inspired by Markov reward processes (MRPs), the values of known grid cells are propagated through the values of unknown grid cells, enabling spatial interactions at arbitrary distances and paths, while enabling specialised local interactions between a cell and all its neighbours. By iterating the core update rule, this system will converge to a stable state where all values have been interpolated.

We proposed two variants of VPint in Section 3.4: SD-MRP, which requires no additional data and propagates values at a static weight (discount rate) at every cell, and WP-MRP, which leverages additional datasets of covariates that are known for the full grid, enabling the use of predicted, locally specialised spatial weights between neighbouring cells. We found VPint, especially WP-MRP, to be a method that effectively interpolates missing spatial data, that satisfies the requirement of retaining both local and global spatial properties. In our experiments, VPint performed better than baseline methods representing both powerful interpolation approaches (Gaussian processes) and advanced machine learning-based approaches incorporating spatial statistics and automated machine learning techniques. VPint also converged to a stable state in relatively few iterations (around 20), and WP-MRP needed only a modest correlation (about 0.1) between the features and target variables to reliably perform better than SD-MRP. *In conclusion, for RQ1, we can effectively interpolate spatial data such that both local and global spatial properties are retained through our proposed method, VPint.*

Limitations: We found that VPint did not yet generalise well to spatio-temporal interpolation problems. In spatio-temporal settings the method, in its current form, should be applied independently for every time step. We also found the performance of VPint to be less favourable for spatially clustered missing data than for randomly missing data, which hampers its application to, e.g., cloud removal tasks in EO data. We, therefore, recommend VPint for problems where data gaps are more uniformly distributed over the space, such as sensor network data or data with a low sampling rate (few measurement points over a long spatial distance).

7.1.2. RQ2: EO DATA INTERPOLATION

The second research question we will answer is RQ2: *How can we effectively and easily interpolate unpredictable, spatially clustered missing data in Earth observation imagery?* (Chapter 4)

For this research question we focused on the EO input data that is also affected by Challenge 1. Although our focus was on missing data due to cloud cover, the methods proposed for this problem would generalise to any type of missing data in EO imagery that would share many of the same challenges, such as gaps introduced by non-overlapping spatial coverage or Landsat ETM+ SLC-off data gaps. Existing state-of-the-art cloud removal methods often come in the form of specialised deep neural networks, which show impressive numerical performance on curated datasets, but are difficult to apply in practice and typically need to be re-trained for every type of sensor, band subset, resolution and type of missing data. The resulting limited practical uptake means that, most of the time, cloud removal is performed by mosaicking cloud-free pixels from older observations into the cloudy pixels of new images. Such approaches are convenient, but have substantial drawbacks in terms of numerical performance.

VPint, presented in Chapter 3, was a good candidate as an alternative cloud removal method with the same requirements as mosaicking approaches, while potentially offering far greater numerical performance. The existing VPint algorithm, however, could not be applied effectively to this new problem setting without substantial modifications and extensions.

In our proposed VPint2 algorithm, we kept the core concepts of VPint interpolation, but modified WP-MRP to compute highly precise spatial weights from a reference image, instead of estimating them from correlated covariates. This weight computation approach, which is available only in special cases, such as EO data and similar spatio-temporal settings, where regular measurements of the exact same spatial area are taken, enabled the application of VPint2 to image processing tasks. In addition to this modification, we introduced identity priority and elastic band resistance to the VPint2 update rule. These extensions greatly improved the stability and performance of the method for EO data, as they limit the impact of spatial relationships that are likely to have changed between the reference image and the target cloudy image.

Through our experiments, we found VPint2 to perform significantly better than competing methods, including specialised deep neural networks, in 17 out of 20 conditions spanning diverse geographical locations, land cover classes and temporal distances of the reference image. Unlike the original VPint algorithm, which performed notably worse for larger, spatially clustered missing data, VPint2 was not strongly affected by the size of the clouds being removed. The temporal distance of the reference image had a modest impact on the performance of VPint2

and the competing methods. We further found that many of the cloud removal methods showed a particularly strong performance in different parts of the images, and an oracle experiment demonstrated great potential for improvement in cloud removal performance if an effective ensembling method could be found. *In conclusion, for RQ2, we can effectively and easily interpolate unpredictable, spatially clustered missing data in Earth observation imagery using our proposed method, VPint2.*

Limitations: The current VPint2 method relies on a single, completely cloud-free reference image. In some applications, such an image may be difficult to obtain, and a user may instead have access to a time-series of cloudy images. An adaptation of VPint2 to derive weights from the cloud-free parts of the images in this time-series may be valuable. Additionally, while our oracle-based experiments showed the promise of ensembling-based approaches, our work did not include a practical implementation of such an approach, which may improve performance beyond that of VPint2 in its current form.

7.1.3. RQ3: PARAMETER ESTIMATION ILL-POSEDNESS

The third research question we will answer is RQ3: *What makes parameter estimation an ill-posed problem, and which factors affect the reliability of parameter estimation results?* (Chapter 5)

For this research question, we were interested in how Challenge 2, namely noise and ill-posedness, affected the reliability of the solutions of parameter estimation. In Chapter 5, we focussed on the inversion of the PROSAIL RTM, which we used as a prominent example of a parameter estimation method using EO data. The inversion of PROSAIL is generally known as an ill-posed problem, which is considered a downside of the method; however, we were interested in whether this ill-posedness was a property of the PROSAIL inversion specifically (where a single, best-fitting configuration must be found for an observed spectrum), or whether the parameter estimation problem itself (where the real-world parameter configuration must be estimated from noisy spectral observations) was the source of ill-posedness. We were unable to find any existing systematic analyses of the problem that could provide evidence for the ill-posedness of PROSAIL inversion specifically in cases where the parameter estimation problem itself was well-posed. This raised some doubts on whether the challenges encountered by users attempting PROSAIL inversion to perform parameter estimation, where a unique solution could often not be found, were truly caused by the ill-posedness of RTM (PROSAIL) inversion, or rather a general property of the parameter estimation problem.

In Chapter 5, we performed a systematic analysis, testing for the ill-posedness of RTM inversion (specifically the PROSAIL vegetation model), alternative poten-

tial causes for effects commonly ascribed to RTM inversion ill-posedness, and how adding prior knowledge can alleviate the problems. Our empirical results indicated that, unlike what is often assumed, the RTM inversion met all the requirements of a well-posed problem. This suggests that the unreliable results obtained by users were caused not by limitations of the RTM inversion approach, but rather by inherent properties of the parameter estimation problem. We hypothesised two possible properties that may cause ill-posedness on the parameter estimation problem as a whole: noise on the spectral observations and spectral mixing. We found that both factors resulted in ill-posedness on the parameter estimation problem, despite the RTM inversion procedure correctly identifying the correct solution to the (flawed) input data it received. Finally, we found that, under these conditions, adding prior knowledge to constrain the search space is an effective method for reducing the ill-posedness of the problem, even if the ill-posedness is a part of the parameter estimation problem rather than the RTM inversion procedure.

These findings indicate that the parameter estimation problem will still be ill-posed, even if other prediction methods (e.g., hybrid models or fully data-driven models) are used. However, because such methods will always provide an answer for the input, and because statistical biases in the model may partially compensate for the ill-posedness, this ill-posedness will be more difficult to diagnose compared to RTM inversion based on numerical optimisation. Our findings point to data-centric approaches, aimed at reducing spectral noise or adding additional data sources, as a promising direction for improving parameter estimation performance and reliability. *In conclusion, for RQ3, parameter estimation is likely an ill-posed problem due to flawed input data, with factors such as data noise and spectral mixing affecting the reliability of parameter estimation results.*

Limitations: With the exception of our experiments on the loss landscape properties of PROSAIL inversion, our analyses were carried out primarily on simulated data. This was a deliberate choice, because it enabled a highly accurate evaluation approach that would not have been possible with flawed ground truth data from real-world settings (see Section 2.1.3 for why such data is not suitable for accurate evaluation). In particular, given the subtlety of the differences in the loss function landscape that results in ill-posedness for small amounts of observation noise, even small inaccuracies in the ground truth data would have made these analyses challenging to carry out. However, without a full empirical study on real-world data, we cannot know with certainty that data-driven approaches trained fully on real-world data would be affected by the same factors as RTM inversion on simulated data, despite our results strongly pointing to such an effect.

7.1.4. RQ4: POSSIBLE SOLUTION SET

The final research question we will answer is RQ4: *How can we automatically extract the set of possible solutions to a noisy inference problem?* (Chapter 6)

For this research question we were interested in the severity of the ill-posedness caused by Challenge 2, through the factors identified in RQ3. We generalised the findings from parameter estimation to the general problem class of noisy inference and model inversion problems, of which parameter estimation is an example. Although we have established that noise on the observations (EO data) can cause ill-posedness for an inference (parameter estimation) task, it is important to be able to establish the severity of this ill-posedness for specific problem instances. If the severity is low, the inference results can still be reliable, while high severity greatly reduces the reliability of any solution to an inference problem.

In Chapter 6, we proposed the concept of ϵ-manifolds. These manifolds contain all the potential solutions to an inference problem that could reasonably explain the observations. When the observations are noisy, an incorrect solution will have a lower loss function value than the correct solution. Therefore, any solution that explains the observations, up to a factor ϵ, should be considered a viable solution to the inference problem. Here we made two key assumptions: loss-dependent shifts and loss-dependent origins, where the impact of observation noise is assumed to be governed by the loss function landscape. This focus on the loss function landscape differentiates ϵ-manifolds from statistical distributions, because the (prior) probability is not relevant when we aim to identify all the solutions that *could* explain the observations, not those that *did*. As a result, ϵ-manifolds enable novel types of analyses, including analysing the ill-posedness of parameter estimation problems.

We proposed a novel method, eMMI, to automatically approximate the ϵ-manifold for inference problems. This method was based on constrained diversity optimisation, leveraging different heuristics through its objectives and constraints to efficiently explore the space around the point prediction. Through our empirical experiments, we found that ϵ-manifolds offer substantial advantages over statistical interpretations of the problem, and that eMMI approximated these ϵ-manifolds better than the confidence intervals of baseline methods such as Gaussian processes, Bayesian neural networks and approximate Bayesian computation. *In conclusion, for RQ4, we can automatically extract the set of possible solutions to a noisy inference problem through the approximation of ϵ-manifolds by our proposed method, eMMI.*

Limitations: Although we introduced a theoretical framework for ϵ-manifolds that can be applied to multimodal problems (using ϵ-manifold sets) that may be highly ill-posed, non-convex or chaotic, we tested on unimodal problems, and the heuristics implemented for the current version of eMMI are based on (weak) as-

sumptions of local convexity. Extending the implementation of eMMI to approximate ϵ-manifolds in all possible scenarios, though possible within our proposed framework, was beyond the scope of the work. Instead, we focussed on a thorough analysis for scenarios that were relevant for our objective of physical parameter estimation. We also assumed a static ϵ for all instances of a particular simulator, while in reality, this assumption may not always hold, necessitating an extension incorporating dynamic settings for ϵ.

7.2. OUTLOOK AND FUTURE WORK

Throughout this dissertation we have introduced four concrete contributions, focusing on different parts of the pipeline shown in Figure 1.1, to improve parameter estimation. However, in all these components, there are remaining challenges to overcome, and further opportunities to explore. We start this section by describing two research directions we briefly explored, but ultimately did not pursue. We provide our experience in the hope that other researchers, if they are interested in this topic, will be aware of the hurdles to these approaches, and the conditions necessary for them to become feasible in the future. After this, we will provide our concrete recommendations for the next steps of research that could build on the work contained in this dissertation.

7.2.1. ALTERNATIVE RESEARCH DIRECTIONS

During the research work contained in this dissertation, there were two research directions whose merits we partially explored, but which we ultimately did not pursue. In this section, we will briefly explain our motivation for exploring them, the barriers to a successful execution that we encountered, and the necessary conditions that, in our view, could render them feasible in the future.

AUTOMATED INSTANCE GENERATION FOR SPECIALIST INVERTED SIMULATORS

When examining the most successful examples of model inversion-based parameter estimation, a common theme is that the scope of the study area is narrow and very clearly defined. For example, the model may be used for specific crops, such as maize [55, 52], wheat [46, 54] or rice [53], or specific locations [54, 45]. This enables the use of strict range constraint priors when performing model inversion, as described in Chapter 5, or the biased training of specialised machine learning models whose training data only includes solutions relevant to the study area. Our objective was to perform parameter estimation in general problem settings: the method should be applicable to any location and any species, or indeed, any type of physical simulation model.

We considered large, generalist models to be unlikely to be effective, because

of the ill-posedness of parameter estimation. Therefore, we were interested in borrowing concepts from the AutoML community in benchmark instance generation [293] to automatically generate specialised LUTs for local areas. We intended to match the distribution of the simulated spectra in the LUT to the distribution of the observed spectra for the local study area. The principle behind this approach was that we aimed to avoid a situation where a model would need to predict one value for one context, and predict another value for the same input data in a different context (e.g., a different geographical area or season). The main obstacle to this approach was obtaining ground truth data for the simulated data. Although we could perform post-hoc selection of instances using the simulated spectra, these matching spectra could still have been generated from very different sets of parameters. Therefore, while we maintain that the automated training of specialist models may be an effective method for global parameter estimation with minimal assumptions, we concluded that the implementation of such a method would be infeasible until the ill-posedness of model inversion approaches was addressed. Although we explored this ill-posedness in Chapters 5 and 6 by identifying its sources and enabling analyses, further work is still needed to truly overcome ill-posedness, enabling a resumption of this line of research.

PHYSICAL CONSTRAINTS AND EQUATION DISCOVERY

In the real world, physical parameters are often not independent, as these variables affect and are affected by a large, causal system consisting of many physical parameters and their relationships [294]. In simulation models, such as the PROSAIL model we have covered extensively in this dissertation, such relationships are usually not taken into account. For example, in PROSAIL, the physical relationships modelled by the simulator relate to the behaviour of light when encountering various media described by the physical input parameters of the model, such as water- and chlorophyll content. These input parameters can be independently manipulated by the user, and the model will generally perform a valid simulation using these values, regardless of how physically plausible their combination is. In reality, many of these input parameters could affect one another (e.g., a high value for one parameter is only possible if another parameter also has a high value, or they are mutually exclusive).

This led to the idea that capturing such physical dependencies may be a promising approach for imposing constraints on the search space of the inversion of physical models, such as PROSAIL. Imposing such constraints may have resulted in a similar reduction in ill-posedness as could be seen for range constraints in Chapter 5. Although this approach of constraining the search space using codified domain knowledge is conceptually appealing, to our knowledge, so far the relationships between such parameters have not been studied and formalised to an extent that

would be sufficient to construct, e.g., constraints for an optimisation algorithm.

To overcome this problem, we considered using equation discovery algorithms, such as ProGED [295], to automatically extract such relationships from data. However, we found the available data to be insufficient to make this approach feasible. Although isolated, small datasets exist, such as the ANGERS database [296], the available data is currently too limited in scope to reliably extract general constraints applicable to a global scale. Other, larger datasets may themselves consist of estimates that may introduce inaccuracies into the equation discovery procedure (e.g., NEON data [297]), or may not contain measurements of all the relevant parameters. However, should the availability of data improve, or should studies by domain experts result in formalised relationships between input parameters, a constraint-based approach to alleviate ill-posedness may become more viable.

7.2.2. FUTURE WORK

We believe that there is great promise in extending our work in the directions described in the remainder of this chapter. Our recommendations, like our contributions in Chapters 3–6, will consider the parameter estimation problem from the perspectives of the different components of the pipeline shown in Figure 1.1. Together, these directions can be combined to further improve inference performance and reliability, and learn even more about the parameter estimation problem. We will conclude this chapter, and this dissertation, with a hypothesised pipeline of what such a combined setup could look like in the future.

NEURAL VPINT

In Chapter 3, we introduced VPint as an independent spatial interpolation method that can fill in missing data in a target dataset. In the WP-MRP method, the spatial weights of the dataset were computed by a machine learning model based on a feature dataset of covariates, and the system of grid cell values was iteratively updated using these weights. This approach, while effective, required user input in the form of suitable feature datasets and choices on how the spatial weights were computed, while the use cases were fairly specialised (spatial settings where a gridded representation of point data is desired). This limits the applicability of VPint in parameter estimation settings where relevant covariates cannot be measured – for example, many of the PROSAIL parameters described in Chapter 5 are equally difficult to measure.

Moreover, applying VPint independently prior to any downstream deep learning model risks discarding the advantages (particularly with regard to keeping many parameters trainable) that deep learning has to offer. The VPint algorithm lends itself very well to an implementation as a neural network 'VPint block'. When incorporating VPint as a component of a neural network, many of the choices that

currently need manual input could be automatically learned from data, such as the computation of the spatial weights or the masking of missing data. Iterations could be represented by layers, where each cell is connected to its neighbours in the next layer, and the trainable weights of these connections (the spatial weights in VPint) are shared between all layers, boosting efficiency. The weights could be either predicted from covariates via neural network connections, or trained independently from a training dataset of the same area of interest. The network could also, in previous processing steps, learn to identify and mask missing data automatically.

Integrating VPint into deep learning pipelines in this manner, where users only need to specify a 'VPint block' in their network to remove any possible missing data from their input data, could greatly improve its applicability to general image processing tasks, as well as EO data specifically. Moreover, the performance of deep learning models, which represent the state of the art for many problems, may be significantly improved through the inclusion of a dedicated, efficient block to explicitly remove missing data. The research question for this approach would be: *How can VPint be integrated into existing deep learning frameworks such that missing input data can be automatically identified and interpolated in an effective and efficient manner?*

TIME-SERIES VPINT2

When using VPint2 to fill in missing satellite data in particular, as we did in Chapter 4, our proposed method was designed to interpolate missing data (especially clouds) from a single image, and used a single, fully observable reference image to guide the reconstruction. While this setup can be convenient in setups where only a single target image is of interest, other applications may be presented with a time-series of images with potentially missing data, and require the missing data in the full time series to be filled in. In existing work, time-series data already forms an appealing source of information complementarity, exploited by multiple deep learning-based cloud removal methods [41, 172, 174]. It seems probable that VPint2 could likewise be adapted to this problem setting, particularly when combining this approach for VPint2 with the neural network version of the original VPint described above. This would further reduce the data requirements of the method, and potentially enable bulk processing to generate full datasets without missing data.

In a basic version of the algorithm, the spatial weights used by VPint2 could be computed from the observable parts of the images in the time-series, in a manner similar to that of the current VPint2 algorithm. When multiple images have cloud-free data available for the same pairs of pixels, algorithm stability could be improved by computing, e.g., the median of the weights, or a recency bias could

be added. If none of the images contain cloud-free data for certain parts of the image, the algorithm could default to a weight of 1 (identity). The neural network version of the algorithm could entail substantial advantages by enabling complex, non-linear relationships between the flawed input time-series data and the spatial weights, enabling the computation of spatial weights using abstract features from a rich latent space. This approach would also enable the model to learn cloud masks and other data imperfections, reducing the need for explicitly defined 'missing data'. This adaptation of VPint2 may be another valuable contribution to improve the reliability of EO data. The research question for this method would be: *How can missing pixels in a time-series of satellite data be effectively and efficiently interpolated when a fully observable reference image is not available?*

SPECTRAL DENOISING

Moving beyond missing data, much of our focus in Chapters 5 and 6 has been on the ill-posedness caused by observation noise in parameter estimation and other inference problems. When this noise, such as atmospheric interference or spectral mixing, is present in the data, our results in Chapter 5 showed that parameter estimation problems become ill-posed, while our proposed methods in Chapter 6 enabled estimations of the severity of the ill-posedness caused by this noise. Although it is certainly useful to diagnose the problems, these analyses point to spectral denoising as a potentially valuable approach to further improve estimation performance. If this denoising task could be performed perfectly, such that the denoised data exactly matches noise-free simulated data, performing parameter estimation would become a regular, well-posed problem, as we found the inversion of the PROSAIL RTM to be in Chapter 5.

In recent years, there have been considerable advances in deep learning techniques, such as denoising autoencoders [211, 298], and data fusion approaches [40, 41, 81]. For EO data in particular, the spatial and temporal autocorrelation of images may enable such deep learning methods to automatically reduce the level of noise on the spectral data. It is also possible that neural VPint or VPint2 blocks could be used to perform this denoising, by automatically masking out unreliable pixel values, or interpolating data with random data masks to perform a type of implicit regularisation. Performing denoising on EO data to reduce the ill-posedness of parameter estimation would result in the following research question: *How could we effectively remove noise from EO data, such that the denoised data closely resembles the noise-free data as simulated by RTMs?*

LATENT SPACE EMMI

Despite the promise of denoising and further improvements to the VPint algorithm, some inaccuracies will be unavoidable, and the inversion some physical

models (e.g., other RTMs with different properties from PROSAIL) may still be ill-posed even if the data is noise-free. In this case, we still need ϵ-manifolds and eMMI to approximate viable solution sets. The results in Chapter 6 showed that ϵ-manifolds and eMMI are highly effective at approximating the viable solution set of various inference problems, notably including machine learning. By constructing ϵ-manifolds for specific instances, users could, for example, find sets of adversarial examples for machine learning models, analyse the ill-posedness of inference problems, or compare model parameterisations.

For many machine learning EO applications (especially deep learning, which represents a large part of the state-of-the-art methods), there are concerns about generalisability and extrapolation, because the complex Earth system is highly diverse (we explained this in Section 2.1.3). The robustness analyses enabled by eMMI would clearly be beneficial to diagnose generalisably problems. However, most current state-of-the-art machine learning- and deep learning models are far more complex and high-dimensional than even the 100-dimensional linear regression models we considered in our experiments in Chapter 6, both in terms of the number of trainable parameters (for model parameterisation problems) and input parameters (for adversarial examples). This means that the current version of eMMI is ill-suited to approximating ϵ-manifolds for state-of-the-art models we are interested in for EO data and data-driven parameter estimation models.

In order to overcome this obstacle and make eMMI directly applicable not just to approximating ϵ-manifolds of RTM inversion problems, but also the analysis of deep learning model parameterisations or their robustness to certain types of input, we would need to greatly improve the applicability of eMMI to high-dimensional parameter spaces. To achieve this, it may be possible to approximate ϵ-manifolds by optimising in a latent space [299, 300], compressing the information contained in the high-dimensional problems associated with large models, thereby enabling ϵ-manifold-based analyses in more practical machine learning- and deep learning settings. Since the loss landscapes associated with such latent spaces would likely be much less favourable than those considered in Chapter 6, this approach may require some extensions suggested in the future work of Section 6.8, such as a surrogate model-based version of the method, and better support for multi-modal, highly non-convex landscapes. The research question associated with this direction would be: *How can we approximate ϵ-manifolds in nonlinear latent spaces for high-dimensional inference problems?*

PHYSICS-INFORMED REPRESENTATION LEARNING

Finally, we recommend approaching parameter estimation from EO data from another angle altogether. Although we have mainly focused on addressing the challenges of parameter estimation through AI techniques, there may be great promise

in doing the opposite: addressing AI challenges through parameter estimation, and notably, RTM inversion. In parameter estimation, we often use a form of hybrid models: machine learning-based models whose training was supervised using a training dataset generated by an RTM (usually PROSAIL). However, there are other types of hybrid models, most notably including physics-informed neural networks (PINNs) [33, 36]. These models can take multiple different forms; the form most relevant to us consists of a standard neural network encoder, whose latent representation forms the inputs of a physical model based on differential equations. The loss of the final output prediction by the physical model can be propagated back, because its differential equations lend themselves well to back-propagation.

This type of setup could be replicated by an RTM, where a neural network takes EO data as input, and encodes this to a latent representation that forms the input for an RTM. The RTM can then be used to simulate EO data based on these inputs, which should match the input spectra provided to the neural network. This type of physics-informed neural network has recently been successfully applied to PROSAIL inversion [301], showing the promise of this research direction. The neural network could learn a mapping from EO data to RTM input parameters, without needing any ground truth dataset containing these parameters. If desired, given that this approach would likely still suffer from the same ill-posedness as all other parameter estimation methods, the model could also learn to predict a diverse set of solutions in the ϵ-manifold, obtained by eMMI, and possibly process these solutions further into a latent representation, to summarise the ϵ-manifold. This approach would address several key limitations of conventional, fully data-driven approaches (see Section 2.1.3), by eliminating the need for accurate ground truth measurements and sufficient support (representation in the training dataset) for all possible viable solutions.

In addition to direct parameter estimation or the amortised estimation of the associated ϵ-manifolds, existing representation learning techniques (e.g., for foundation models) could be supplemented by a this latent representation of the ϵ-manifold for PROSAIL inversion, thereby guaranteeing the inclusion of a set of informative features highly relevant to known physical processes related to, for example, vegetation. This would lead to the following research question: *How can we train a model to learn a representation of physically relevant features for EO data, in a fully unsupervised manner?*

IDEALISED FUTURE PIPELINE

Based on the recommendations above, an idealised pipeline (visualised in Figure 7.1) may take the following shape. The user receives a time-series of satellite images as input, with a goal of returning an estimated distribution of parameter val-

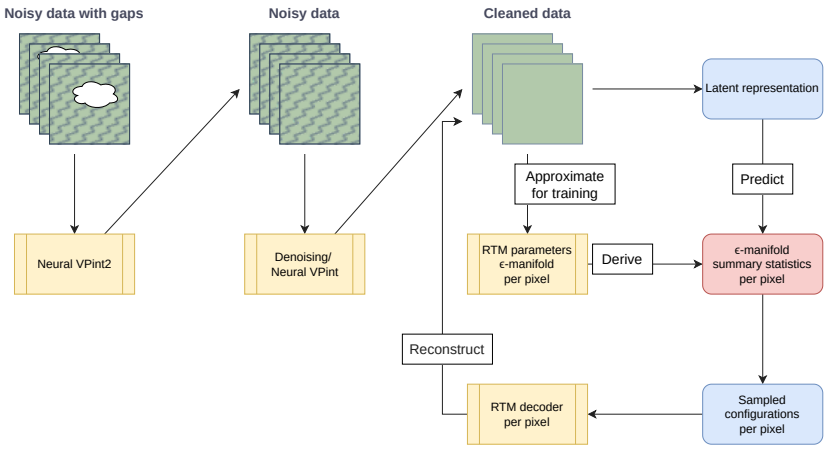

Figure 7.1: Idealised future parameter estimation pipeline.

ues at every pixel. We use a deep neural network to enable efficient large-scale processing and take advantage of their state-of-the-art performance, particularly for problems with large volumes of data available (e.g., EO data), and train this network as follows.

i) All missing data, such as clouds or faulty pixels, are first processed using a time-series neural VPint2 block. ii) The user can choose to use the previous VPint2 block to simultaneously perform denoising, or use a separate neural VPint block for this purpose. iii) The model next maps the time-series data to a latent representation using standard deep learning layers. iv) For all valid pixels in the denoised input, we use eMMI (possibly its latent version) to estimate the ϵ-manifold for every pixel in the data (this will be computationally expensive at training time). v) We estimate summary statistics for the ϵ-manifold from the latent representation of iii) in a manner similar to amortised Bayesian inference (this will be a lossy conversion, because ϵ-manifolds are not themselves a statistical distribution). vi) Based on the estimated statistical distribution, we sample from this distribution to pass inputs to an RTM, which reconstructs pixel values. vii) We train the network on the either a regression loss (e.g., mean squared error) for the pixel values throughout the entire time-series, or a pixel-wise classification (segmentation) loss measuring whether the spectra of reconstructed pixels are within a value of ϵ of the original spectra.

Of course, the above pipeline is highly idealised, and assumes that every step

of this procedure functions as intended. Many of these steps could likely become new research projects on their own. However, if successful, the training pipeline described above would result in a trained model that a) can perform large-scale predictions for parameter estimation, b) automatically fills in data gaps and denoises the data, c) requires no in-situ ground truth data to train, and d) incorporates ill-posedness (including for unobserved configurations, which would be a major limitation of methods not using eMMI, given the generalisation concerns described above) in the uncertainty of its predictions.

Although this idealised pipeline is still many years of intensive research work removed from being realised, we consider the recommendations we described in this chapter to be the most promising concrete next steps to take to bring it closer to fruition. It is our hope that, one day, we could see a parameter estimation approach similar to what we described above be implemented and deployed in practice.

7.3. CONCLUDING REMARKS

Physical parameters describing the current state of the Earth, such as ecological parameters describing vegetation [191], atmospheric parameters describing gases and particles in the air [1, 2] and marine parameters describing our oceans [3], are of tremendous importance to scientific research and informed decision making in, for example, environmental policy and public health campaigns. We would like to estimate these parameters indirectly from the abundant Earth observation data collected by satellite platforms, thereby creating regular, global maps of important environmental and scientific parameters. Unfortunately, performing this estimation is greatly hampered by data gaps, noise and ill-posedness, resulting in unreliable and inaccurate estimations.

In this dissertation, we have proposed multiple methods for increasing data reliability and diagnosing ill-posedness. We have also identified observation noise as the likely primary cause of ill-posedness for parameter estimation problems, leading to concrete recommendations for future work to focus on improving the data quality and information content in the observed data. We believe that these findings are highly relevant to domain practitioners, who depend on accurate parameter estimation results to perform their research, but may misattribute inaccurate results to the methods being deployed, rather than a fundamental characteristic of the problem. While particularly severe in parameter estimation, data gaps, noise and ill-posedness are not unique to this problem. We invite any researchers from different fields, if they are interested in the solutions presented in this work, to apply them to their own problem settings.

Although, befitting scientific research work, we wrap up this dissertation with

more questions than we had before we started, it is our hope that the findings and methods contained in this dissertation contribute to the collective scientific efforts across disciplines to better understand, monitor and maintain the state of the Earth.