Advancing Physical and Stochastic Methods in Reactive Transport Modeling of the Hyporheic Zone – Improved Biophysics and Deep-Learning-Based Upscaling

Abstract

Within the hyporheic zone, a complex interplay of abiotic processes dictates the growth conditions of biomass. Factors like soil heterogeneity, stream conditions, temperature fluctuations, and nutrient availability all converge to shape this environment. Given the hyporheic zone's potential role to bioremediate contaminants through biotic and abiotic reduction, decoding these growth determinants has broader ecological significance. In this study, we present a Monte-Carlo-style exploration into how varied initial conditions influence biomass growth. Our modeling approach simulates a two-dimensional, meter scale cross-section of the hyporheic zone, integrating accurate permeabilities and hyporheic flux, and modeling chromium reduction through Monod kinetics.

To effectively capture bioclogging dynamics and soil respiration, we've enhanced the reactive transport model, PFLOTRAN. Our expanded model accounts for biomass decay influenced by fluid velocity, and the dependency of biomass growth on temperature. Our diverse simulation conditions, augmented by sensitivity analyses for pivotal abiotic factors, offer holistic insights into microbial growth dynamics through mean trend analysis, mean spatial distribution analysis, PCA and clustering, and correlation heatmaps. This analysis reveals a number of unreported phenomena, and shows that abiotic reduction is more dominant than biotic reduction, even in cases of high biomass concentrations. Furthermore, our analysis shows that flow conditions and permeability heterogeneities have a less significant impact on biomass growth than is commonly expected.

In addition to our physical enhancements to PFLOTRAN and our investigation of correlated features in our simulations, we also present a deep-learning-based upscaling model. We test two different amounts of upscaling (5x and 10x), and show that our method offers high accuracy, enabling safe upscaling that can dramatically speed up Monte-Carlo-style simulations. We further show that upscaling can be performed on features that were not trained on, indicating the general ability of our model to upscale most features in a reactive transport simulation.

Our work seeks to improve both improve understanding and provide a useful tool for the research community. We investigate the nuanced relationships between abiotic elements and bacterial proliferation and highlight how these relationships impact chromium reduction in the hyporheic zone. Furthermore, we provide an upscaling tool to the broader research community that can be used to decrease the run times of high-resolution reactive transport simulations. Through this multi-faceted study, we offer a fresh perspective on the modeling of reactive transport in the hyporheic zone.

1 Introduction

In the vast realm of environmental science, the hyporheic zone stands out as a complex interface that has captured the attention of researchers for decades (**cite**). This subsurface region, generally defined as the interface between river water and groundwater (**cite**), hosts a myriad of complex interactions (**cite**), with biomass serving as a central character influencing broader hydrological and biogeochemical cycles (**cite**).

Historically, studies in this domain have primarily revolved around understanding the physical properties of the hyporheic zone, such as its flow dynamics and permeability (**cite**). However, recent research has started to illuminate the critical role of biofilms and their influence on these physical properties (**cite**). The emergence of biofilms and their subsequent decay—collectively known as bioclogging—have been recognized as significant factors affecting water flow and solute transport (**cite**). Furthermore, the application of advanced computational tools has provided deeper insights into these intricate dynamics, revealing a delicate balance of feedback mechanisms that drive biomass growth (**cite**).

Biomass growth in the hyporheic zone can theoretically be considered a giant feedback group of a myriad of processes. Flow, as an initial causal variable in most cases, acts as the foundation for this feedback loop. It instigates a flux of temperatures and nutrient concentrations within the hyporheic zone. This flux, in turn, triggers changes in biomass concentrations. As a general observation, higher nutrient levels and increased temperatures tend to promote growth (**cite**). However, the relationship between temperature and biomass is multifaceted. For instance, an uptick in temperature leads to reduced water viscosity, which subsequently results in increased flow speeds. Beyond a certain critical shear threshold in our model, heightened flow velocities can actually hinder biomass, creating a self-regulating balance in biomass growth. This balance, while stabilizing mean concentrations, might introduce greater spatial variability in biomass distributions. Consequently, the role of temperature in biomass dynamics is twofold: directly influencing growth and indirectly affecting flow characteristics.

Biomass growth can also influence a soil's permeability (**cite**). As biofilms become denser, they lead to bioclogging, reducing soil permeability. This impedes the nutrient dispersion across the hyporheic zone, subsequently slowing biomass growth. This dynamic between biomass and permeability creates a negative feedback loop within the overarching feedback mechanisms. More biomass leads to reduced permeability, which in turn diminishes biomass growth. Remediation of bioclogging in our simulations arises from natural decay over time and high-velocity flows that erode the bioclogged materials.

Reactive transport (RT) simulations have emerged as powerful tools that allow for investigation and prediction of phenomena within the hyporheic zone (**cite**). However, large scale RT simulations, and Monte-Carlo-type investigations, can take a large amount of time to complete, since these simulations require large numerical computations that depend on convergence criteria (**cite**). Furthermore, RT simulations may fail to converge, which may cause problems for sensitivity analyses.

In this paper, we aim to provide two major contributions. First, we seek to further our understanding of the hyporheic zone by investigating the complex relationships governing biomass growth for a variety of input conditions. Furthermore, we improve the physical accuracy of biomass growth in PFLOTRAN by making it a function of temperature and flow speed. Second, we provide a deep-learning-based method for the upscaling of RT simulations, which allows for a 30x speedup in the generation of large-scale simulations.

2 Methods

2.1 Model Formulation

2.1.1 Description of Chrotran

Our simulations are based in PFLOTRAN, a multi-physics reactive transport simulator developed by multiple national labs (**cite**). Specifically, we have adapted the Chrotran (**cite**) version of PFLOTRAN to represent a high complexity simulation of biomass growth in the hyporheic zone at the Darcy scale. Chrotran defines biomass growth as a function of electron donor (ED) concentration through simple Monod kinetics. It uses biotic and abiotic reactions to model Cr(VI) reduction, defines a mobile-immobile mass transfer system for biomass and ED, and allows for bioclogging modeling capabilities via the dependence of porosity and permeability on biomass concentration. A full description of Chrotran can be found at (**cite**).

2.1.2 Augmentations to Chrotran

The first augmentations we made to the published version of Chrotran were to the given parameters. Specifically, we matched our biomass growth rates to the small amount of data that has been collected on biomass growth in the hyporheic zone (**cite**), and we matched the steady-state concentrations to be generally representative of those found in wetlands (**cite**). In the published version of Chotran, biomass decay is defined by a simple linear decay function that only considers a minimum biomass and the current amount of biomass. However, recent research has shown that biomass decay can also be thought of as a function of shear stress (**cite**). At sufficiently high flow speeds, large shear stresses forming at permeability boundaries may be great enough to dislodge and advect biofilms and other bacterial deposits. Although we would ideally use shear stress values to alter the biomass decay function in Chrotran, we lack the capabilities to calculate shear stresses within PFLOTRAN. Instead, we calculate the cell-specific biomass decay as a function of cell-specific Darcy velocity **(equation 1)**.

Equation 1: $\lambda_b = \lambda_{B2}(B - B_0)(v - \alpha)^{\beta}$

B2 represents the background decay rate based on environmental factors. B is the concentration of biomass (mol/m3), and B0 is the initial, or background, biomass concentration. V represents the Darcy velocity of the fluid. Alpha is a fitting parameter that physically represents the fluid speed required to start the shearing of biomass. Beta is a fitting parameter that physically represents the initial decay value and the rate of increase of decay as flow speed

increases. The idea here is that initial scouring results in a steady-state amount of biomass at a certain flow speed, then an increase in the flow speed will result in more scouring due to both the increased flow speed and the formation of preferential flow channels in the biofilm. The fitting parameters alpha and beta were tuned through comparing our results with published research on biofilm thickness as a function of shear. We calculated shear during post processing, and compared its spatial variations to spatial variations in velocity and biomass. Where shear was below 1e-2 Pa, we had an insignificant change in biomass due to flow speed, and above this value, the amount of decay scales roughly exponentially with flow speed. Through this method we fit the values of alpha and beta to respectively be 5e4 and 2.2. The fastest speeds in our simulations were about 2.8e-3 m/s. Plugged into equation 1 with our derived values of alpha and beta, this gives us a maximum velocity-based decay parameter of 51,745, meaning the biomass decay at this velocity is about 50,000 times greater than the natural decay rate.

Furthermore, we altered the biomass growth function to depend on temperature. Numerous studies have shown that microbial growth generally increases with increasing temperature (**cite**). To include this dependency in PFLOTRAN, we parameterized the Ratkowski function (**cite**).

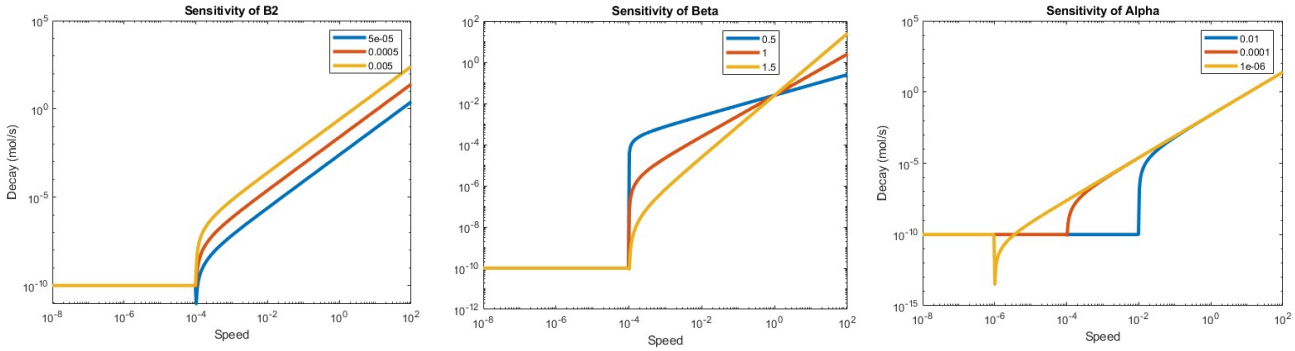

Figure 1. Sensitivity analysis of parameters in the augmented biomass decay equation.

2.1.3 Permeability and Flow

Studies have shown that biomass growth in the hyporheic zone is strongly linked to the hyporheic flux (**cite**). Depending on the concentration of nutrients and the flow speed of the groundwater and surface water flows, a positive or negative flux can have different impacts on growth. To gain deeper insight into how exactly these differences affect biomass growth, we simulated the hyporheic zone under a variety of realistic flow conditions (**Fig. 2a**). The gaining and losing flow conditions at high speed represent the largest hyporheic fluxes we were able to find in the literature (**cite**), the low-speed gaining and losing conditions represent much smaller fluxes, and the medium-speed conditions are an average between the low-speed and high-

speed conditions. The final set of hydrographs we use come from hyporheic flux data measured at the Hanford site (**cite**).

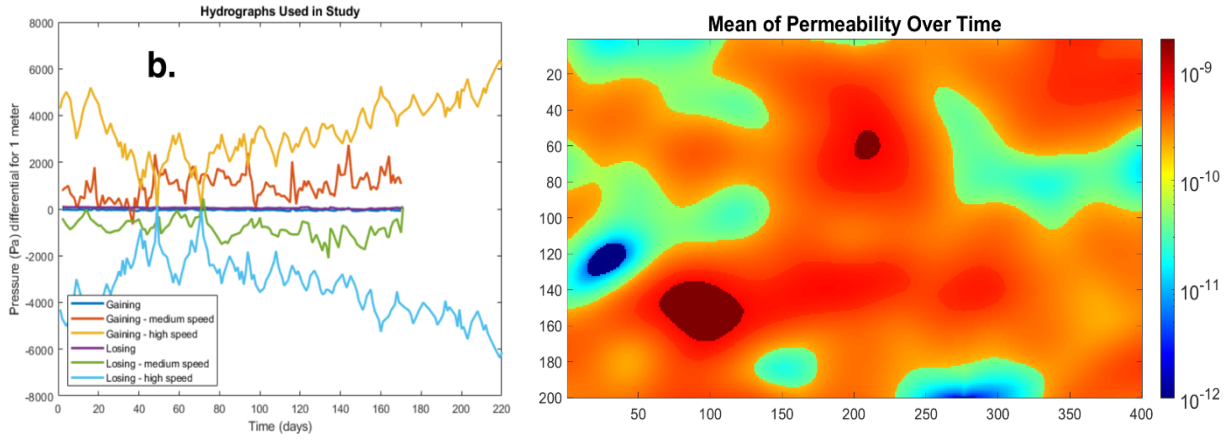

Figure 2. (a) Time-series of pressure boundary condition groups used in the simulations. For each group, slight variations to the time-series were introduced to develop a much wider variety of potential flow conditions for the simulations. **(b)** Example of a heterogeneous permeability field used in our simulations.

In addition to the general direction and magnitude of flow, permeability has a significant influence on the general transport of nutrients and biomass within the hyporheic zone. At low permeabilities, biomass and nutrients are less able to disperse throughout the hyporheic zone, so average biomass concentrations over a large area will be less. However, hotspots of high biomass concentration may still form in low permeability zones, which may also result in bioclogging (**cite**). To understand how exactly biomass growth is impacted by a variety of permeabilities, we created a variety of heterogeneous permeability fields with different covariance ratios (**Fig. 2b**). Although the mean permeability was similar for most simulations (around 2e-10 m2), the effective permeability (Keff) would change based on the covariance ratio. Furthermore, we also ran simulations at extremely low permeability to understand biomass growth in a significantly different environment.

2.1.4 Grid and Boundary Conditions

We simulate a 1 meter (in vertical, or direction of hyporheic flow) by 2 meter longitudinal (in direction of river/groundwater flow) slice of the hyporheic zone. The horizontal and vertical spacing of the grid are both equal to 0.01 meters, meaning the 1x2 meter simulation has 100x200 voxels. The top and bottom boundaries (1m difference) respectively represent the groundwater and surface flow, and the left and right boundaries (2m difference) represent a longitudinal extension of the hyporheic zone. The groundwater and surface flow are regulated by dirichlet boundaries determined by the pressure differentials for the respective simulation type (gaining, losing, or variable). The flow from the left and right is regulated by a dirichlet

pressure boundary of 0, which allows for flow seepage in both directions. The transport boundary conditions are identical to the flow boundary conditions. In this study we simulate multiple scales of the hyporheic zone. The baseline simulations have a scale of 1x2 meters, so we describe the 1x4 meter simulations as “2x” scale, and the 1x20 meter simulations as “10x” scale.

2.1.5 Simulation Variables

The variables (also referred to in this work as features) of the simulations, as well as their range of possible values, is given in **Table 1**. All of the features that depend on the time-evolution of the simulation, such as biomass, chemistry, and flow speed, are referred to as the “physio-chemical features”. The variables that are prescribed at the beginning of the simulation and don’t change value over time are referred to as “input variables”. Each simulation ran for 114 time steps, where each time step corresponds to two days.

2.2 Sensitivity and Correlation Analysis

One of the primary goals of this study is to gain deeper insight on the abiotic controls of biomass growth in the hyporheic zone. To this extent, we employ a variety of sensitivity and correlation analyses to understand how different features may impact growth. We use classical sensitivity analysis, changing one feature and keeping all others equal, to determine the individual impacts that each feature may have on biomass growth. Furthermore, we use a Monte-Carlo-type sensitivity analysis to understand feature relationships at a more global level. Specifically, we ran 375 simulations of the hyporheic zone, each under slightly different input conditions, then used PCA and cluster analysis to understand physio-chemical feature relationships and groupings. In addition, we used a correlation heatmap to identify correlations between all the simulation features (both physio-chemical features and input variables).

2.3 Deep-Learning-Based Upscaling

The other primary goal of our study is to present the deep-learning-based upscaling framework, which we show can be used to dramatically speed up reactive transport workflows. The DL upscaling model consists of 6 linear layers with GELU activation (**Fig. 3**), which allows the model to learn the simple linear relationships across all input dimensions required to increase the size of the final dimension in a way that is consistent with PFLOTRAN. We tested two different amounts of upscaling in this study – 5x and 10x. 5x upscaling corresponds to using the 2x simulation as input to get the 10x simulation as output. 10x upscaling corresponds to using the 1x simulation to get the 10x simulation as output. For a single 1x scale simulation, the shape of the simulation is [t, H, W, F], where t corresponds to the number of time steps (114), H corresponds to the vertical extent (or height) of the hyporheic zone (100 pixels), W corresponds to the horizontal extent (or width) of the hyporheic zone (200 pixels), and F corresponds to the number of physio-chemical features of the simulation. For our upscaling model, we found that performance decreased when using all 9 features, partially due to the large shape of the data requiring a small batch size, so all of our model training and testing was done for individual

features (shape = [t, H, W]). Thus, we created a suite of models to be used together for simulation upscaling – one for Cr(VI), one for molasses-mobile, one for molasses-immobile, and one for biomass. The other physio-chemical parameters of the simulation (Vx, Vy, pressure, porosity and temperature) have a host of both simple and robust methods for upscaling (**cite**), and we were not able to obtain results from our DNN that outperformed these known methods.

We used 60 simulations at each scale to train the 5x and 10x upscaling models, and we used 36 simulations to test our models. For the training simulations, we used all hydrographs besides the Hanford ones. For the testing simulations, we used the Hanford hydrographs. In addition, we changed the range of possible variations in initial conditions for our testing simulations to ensure that good model performance indicates a high level of generalizability. For training, we used the Adam optimizer with an initial learning rate of 2e-4 and an exponential learning rate scheduler with beta=.999. We trained with these parameters for 300 epochs, then performed further fine-tuning with a learning rate of 1e-6 for 20 epochs. In addition to presenting the results for single models, we also use an ensembling method to improve our model generalizability.

3 Results

3.1 General Trends

The general trends of the physio-chemical features are shown through their mean feature plots (**Fig. 3**) and mean spatial distributions (**Fig. 4**). The red dotted lines in each time series plot show the time value of the inflection points for biomass growth (**Fig. 3e**). Average biomass concentrations increase very slowly for the first 20 days, increase rapidly for the next 40 days, then increase at a lower growth rate for the rest of the simulation (**Figs. 3d and 3e**). The first inflection point corresponds to the time when biomass growth starts to dramatically increase. This also represents the point that ED and ED_immobile start to significantly decrease, Cr(VI) starts to increase, and porosity starts to decrease. At the highest amounts of growth (around day 40), biomass growth has a clear correspondence with pressure and Vx **(Figs. 3b and 3a)**. The higher the pressure/Vx, the lower the biomass growth, due to the velocity-based biomass decay that we added to PFLOTRAN. We can also observe a simple linear relationship between biomass and porosity (**Figs. 3c and 3d**). An increase in biomass leads to decrease in porosity, although the slope of the decrease in porosity will be determined by simulation parameters such as biomass density and mean porosity. The second inflection point corresponds to the point where biomass growth starts to slow down. This slowdown is initiated by an increase in pressure and fluid speed (**Figs. 3a and 3c**), but the gradual decrease in growth for the rest of the simulation (**Fig. 3e**) is primarily due biomass crowding, which causes a slowdown in growth as the density of biomass in any given voxel increases. As biomass growth slows down due to crowding, the ED and ED_immobile concentrations increase (**Fig. 3f**), causing a relative decrease in Cr(VI) concentrations, indicating the dominance of abiotic reduction even when biomass concentrations are relatively high.

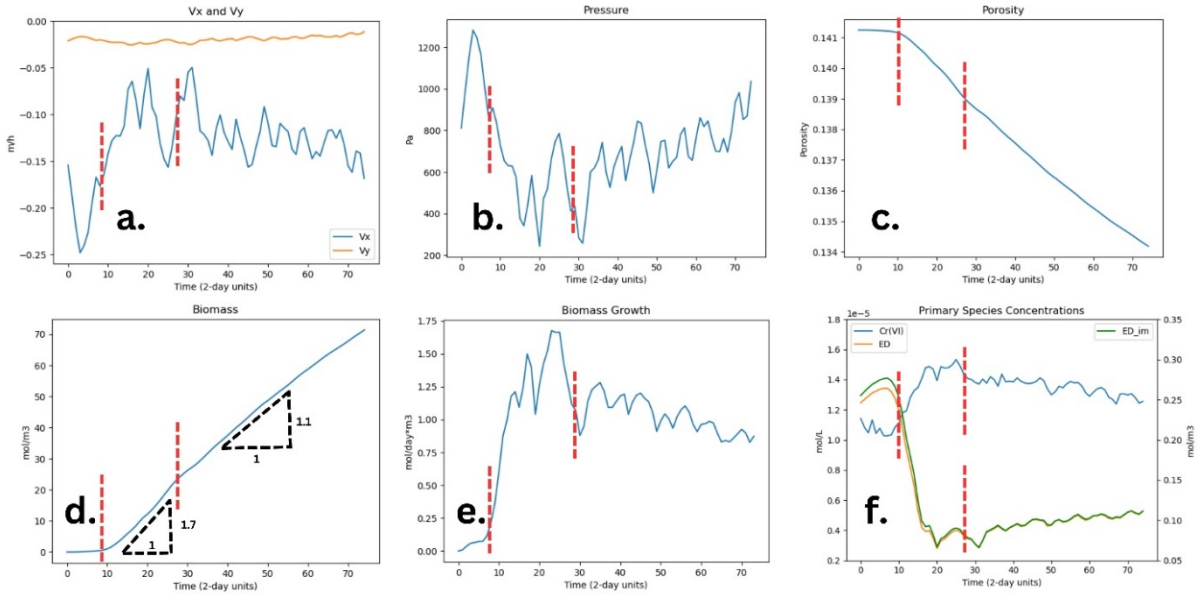

Figure 3. General (averaged over all simulations and spatial dimensions) trends for key physio-chemical features. The red dotted lines show the approximate time of the biomass growth inflection points. **(a)** Vx (blue) and Vy (orange). **(b)** Pressure. **(c)** Porosity. **(d)** Biomass. **(e)** Biomass growth. **(f)** Cr(VI) (blue), ED (orange), and ED_immobile (green). Here, biomass growth is shown to primarily be dependent on ED concentration. The Cr(VI) timeseries shows an increase in concentration as molasses decreases, indicating the dominance of abiotic reduction over biotic reduction.

The mean spatial distributions (**Fig. 4**) further show a high level of spatial correlation between biomass, ED, ED_immobile, and Cr(VI) concentrations. Although higher temperatures lead to greater biomass concentrations due to our inclusion of the Ratkowski equation, we observe very little, if any, spatial correlation between temperature and biomass. Similarly, there is a low amount of spatial correspondence between Vx and biomass (seen in the top left), and Vy and biomass (seen in bottom right), and almost no correspondence between biomass and pressure. There is some correlation between porosity and biomass, and a high amount of correlation between porosity and ED_immobile. ED_immobile represents the concentration of molasses in the immobile phase, and high values correspond to areas with both high total ED and high porosity. When porosity/permeability is high, chemicals will diffuse/advect across permeability discontinuities at higher rates, meaning a greater supply of ED to these zones. In low permeability regions, the biomass is nutrient limited, meaning less biomass growth and thus less molasses is available to sorb. The biomass spatial distribution is almost identical to the ED distribution, except for the slight differences due to Vx, Vy, and porosity. The Cr(VI) spatial distribution is similar to the ED and biomass distributions, but more spread throughout the domain. The similarities are due to the general flow inputs and permeability distributions. Since the groundwater in our simulations contains higher amounts of biomass, ED, and Cr(VI), the section of the hyporheic zone adjacent to the groundwater interface has significantly higher concentrations of these three species. The Cr(VI) distributions are more spread out than those

of ED or biomass because the biomass and ED are not able to completely remediate the Cr(VI) before it reaches the other side of the domain.

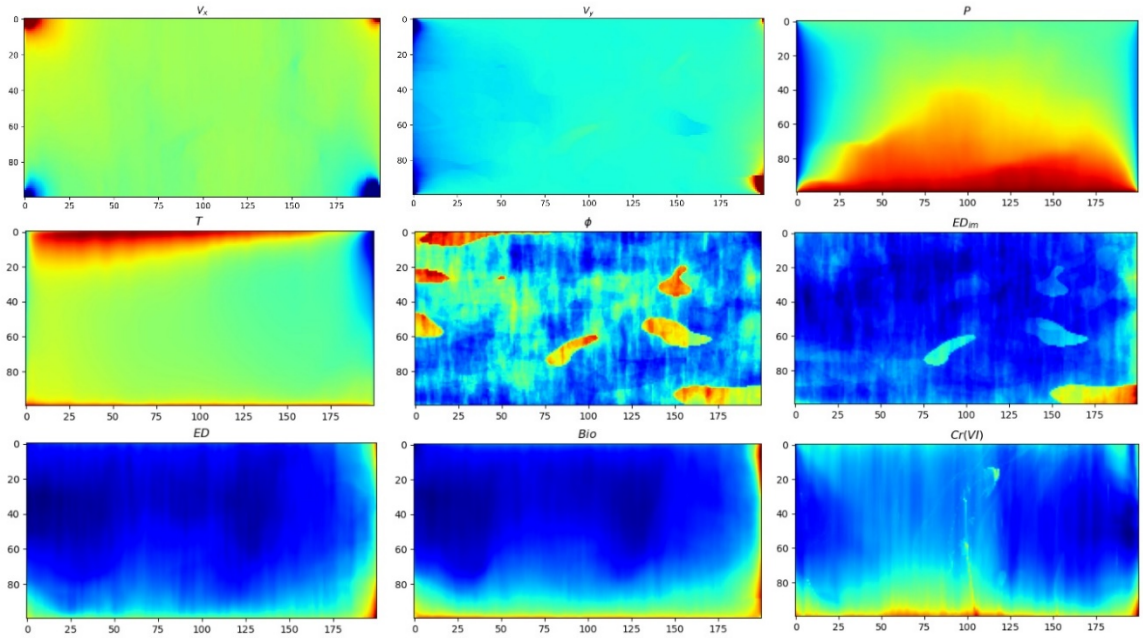

Figure 4. General (averaged over all simulations and time steps) trends for key physio-chemical features. Each variable is normalized, with 0 corresponding to low values and 1 corresponding to high values.

3.2 Sensitivity Analysis

To gain deeper insights into the abiotic determinants of biomass growth in the hyporheic zone, we used sensitivity and correlation analysis. The sensitivity analysis (**Fig. 5**) shows the biomass time series and spatial distributions of the last time step for 6 equally spaced values of a simulation input variable (keeping all other input variables constant). We performed sensitivity analysis for each variable, but only four key variables are presented here, and the rest can be found in **Supplementary Figure 1.** The temperature sensitivity analysis shows that at higher temperatures the biomass growth rate increases, leading to differences in biomass concentrations that remain constant after about day 60. The spatial distributions show more variation in biomass concentrations for higher-temperature simulations. The carbon reuse efficiency (D), is a Chrotran parameter that defines the stoichiometric relationship between the ED and biomass. So for D = 1, one mol of ED creates one mol of biomass. The sensitivity analysis for D shows that lower values of D result in an increase in biomass growth that increases over time. The spatial distributions show that for lower values of D, the biomass is able to spread further throughout the domain. Alpha is another Chrotran parameter that describes biomass crowding. For higher values of alpha, we see slightly lower biomass concentrations over time, although the main difference is in the spatial distributions. When crowding is high (low alpha), we get much higher concentrations of biomass that are

constrained to the first few centimeters of the domain. When crowding is low, we get a lower maximum biomass concentration, but the biomass is spread throughout the entire domain.

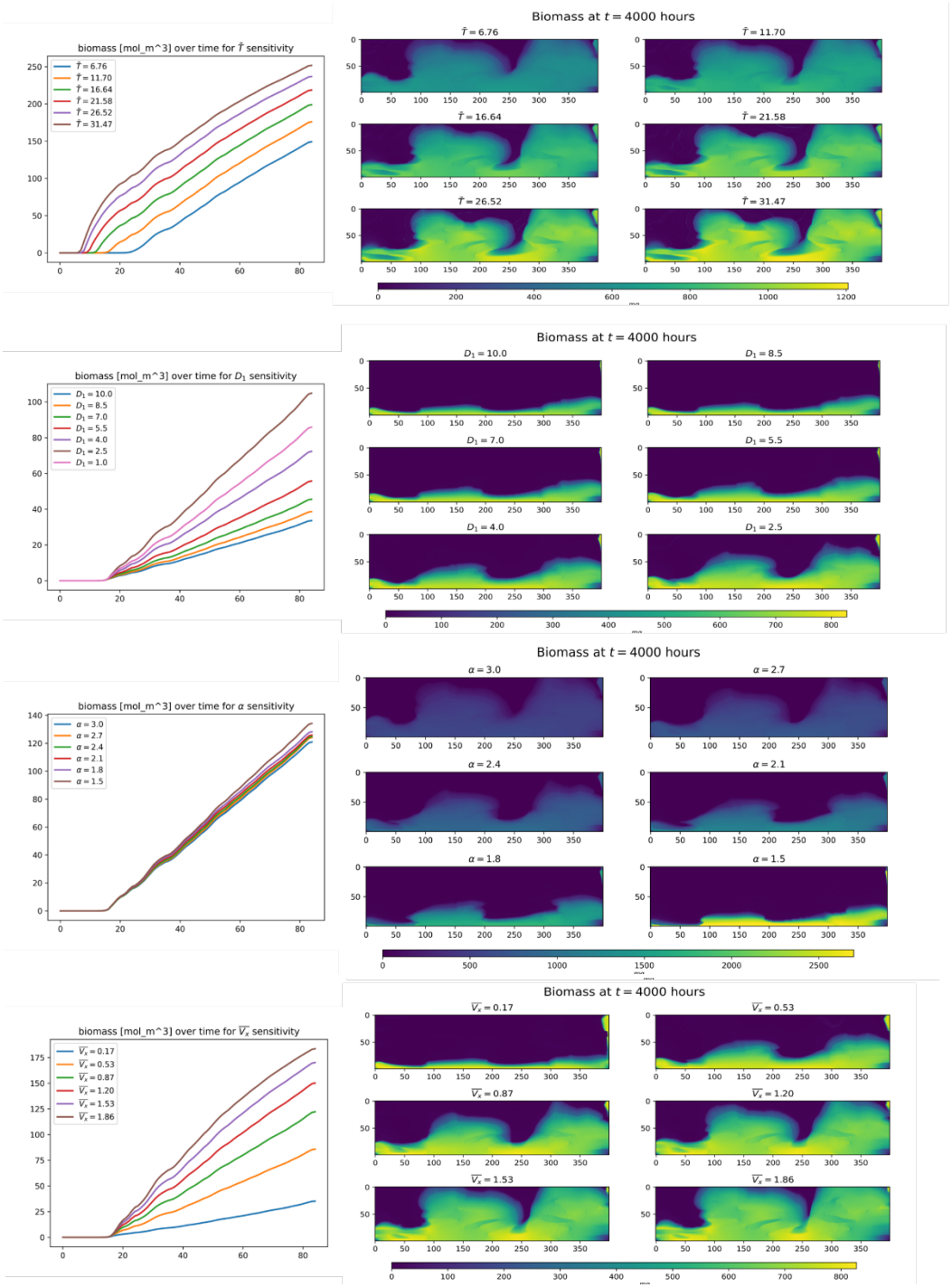

Figure 5. Sensitivity analysis for selected features with large impacts on biomass growth. For each feature, we show the time series, as well as the spatial distributions of biomass for the final time step. The colorbars show biomass concentration in mol/m3. **(a)** Temperature sensitivity. **(b)** Carbon reuse efficiency sensitivity. **(c)** Biomass crowding sensitivity. **(d)** Vx sensitivity.

The Vx sensitivity analysis indicates that greater vertical velocity contributes to a more sigmoidal (as opposed to linear or exponential) growth curve, results in generally greater biomass growth, and causes the growth curve to have small undulations. This wave-like behavior is a result of spikes in flow speed causing significant shearing of biomass, thus briefly decreasing the rate of biomass growth.

To further investigate the interactions between all physio-chemical features, we used PCA and cluster analysis (**Fig. 6**) to identify groupings and large-scale relationships amongst all of our simulation variations. The PCA indicates a significant amount of correlation between ED, ED_immobile, P, σ_P, σ_{Vx}, σ_{Vy}, and $\sigma_{ED\_immobile}$, which are all anticorrelated with T, Vx, and σ_T. This is partially due to the fact that the simulations where the flow was from the river to the groundwater (i.e., losing streams) had the largest vertical velocities and highest temperatures. Similarly, the gaining streams had higher nutrient concentrations, so this grouping likely represents the gaining streams. However, higher pressure is not necessarily correlated with the gaining simulations. Instead, P likely occupies this position in the 2D PCA space because of its high correlation with ED. The PCA shows a slight amount of correlation between biomass and

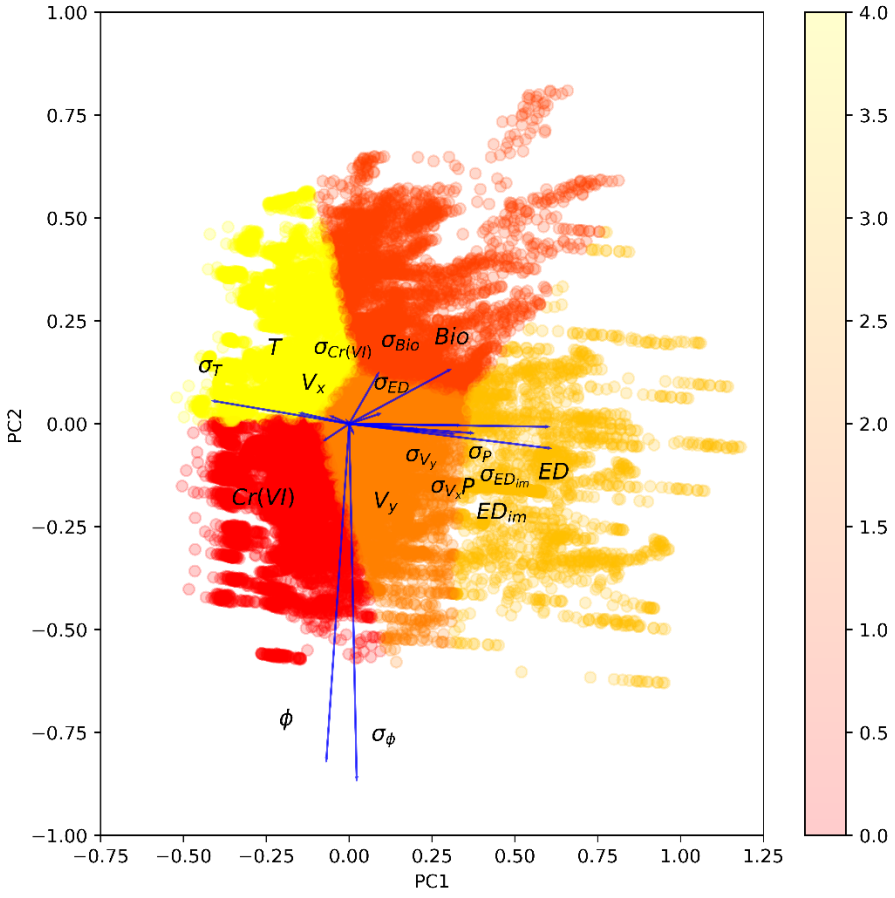

Figure 6. PCA for physio-chemical features that were varied for each simulation. The x-axis shows the first principal component, the y-axis shows the second principal component, and the color bar represents the KMeans clustering output in the 2D PCA space.

ED, although the largest correlations for biomass are with σ_{Bio} and σ_{ED}. Cr(VI) shows complete anticorrelation with biomass, which is somewhat surprising given our other results that show the dominance of abiotic reduction over biotic reduction (**Fig. 3**). σ_ϕ and ϕ aren't strongly correlated with anything, although the show slight positive correlation with Cr(VI) and negative correlation with biomass, indicating that higher porosity soils may have less reductive capacity.

In addition to our PCA of the physio-chemical features, we also use a correlation heatmap to show individual correlations for all simulation variables. Generally the results here confirm the trends observed in the PCA.

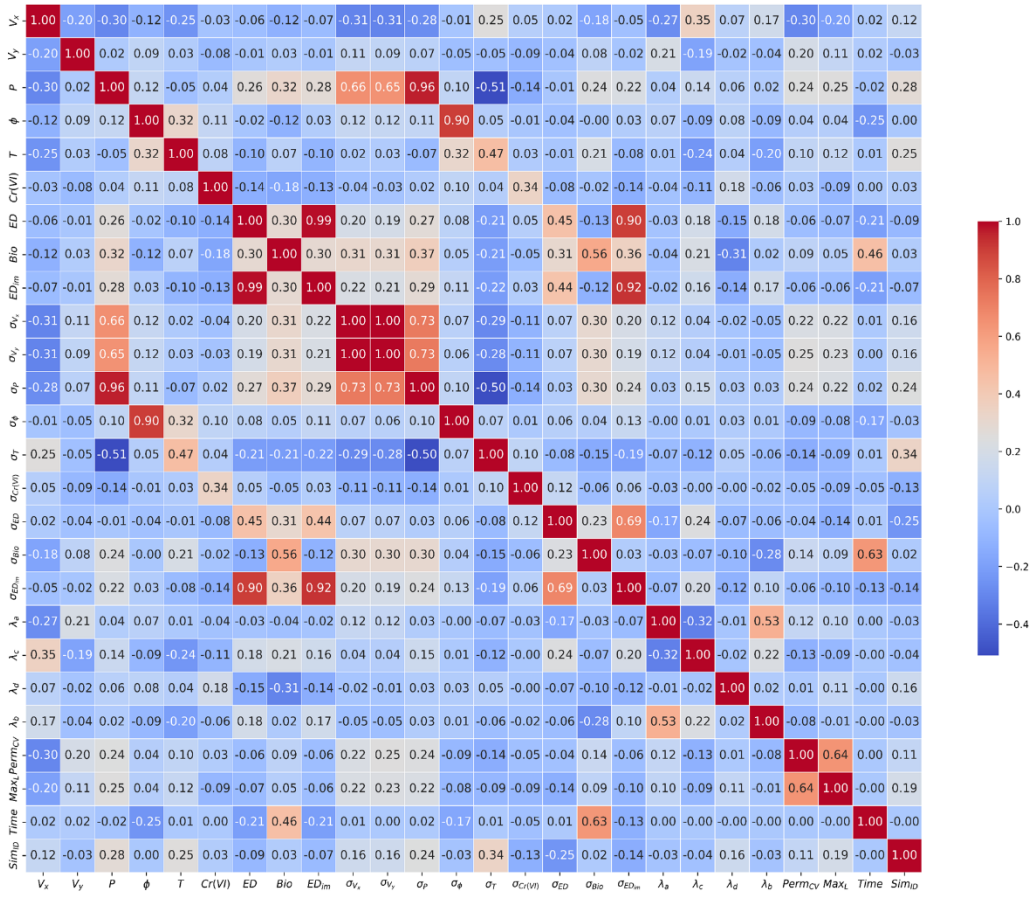

Figure 7. Correlation heatmap for all features that were varied for each simulation.

3.4 DL-Based Upscaling

In addition to our work on the investigation of abiotic factors in biomass growth and chromium reduction, we also provide a DL-based model that can be used to upscale our simulations. RT simulations often take long to converge at large scales, and can sometimes fail to converge, rendering that simulation unusable. To fix both of these issues, we use DL-based upscaling to provide a 30x speedup to the RT workflow. Due to VRAM constraints, each feature was upscaled separately. To test the relative accuracy of our model outputs, we plot the output time series, the input time series, and the ground truth time series (**Fig. 8**). Here the ground truth

time series represents the time series from the 10x scale PFLOTRAN simulation, and the input represents the time series from the 2x scale simulation. For biomass, we find that the upscaling can produce highly accurate time series for 5x upscaling, and somewhat accurate time series for 10x upscaling (**Fig. 8**). However, through model ensembling (training 3 slightly different model variations for each level of upscaling, then averaging their predictions), we were able to improve the generalization of both the 5x and 10x upscaling. The 5x upscaling method provides an average speed increase of 8x, while the 10x upscaling method provides an average speed increase of 30x. Thus, depending on the intended application, we provide researchers with a means to do highly accurate upscaling with 10x time savings, or accurate upscaling with 30x time savings. Furthermore, by providing two fully trained models for different levels of upscaling, we allow researchers multiple ways to produce the full-scale results, in case one or more of the simulations (between 1x, 2x and 10x) doesn't converge.

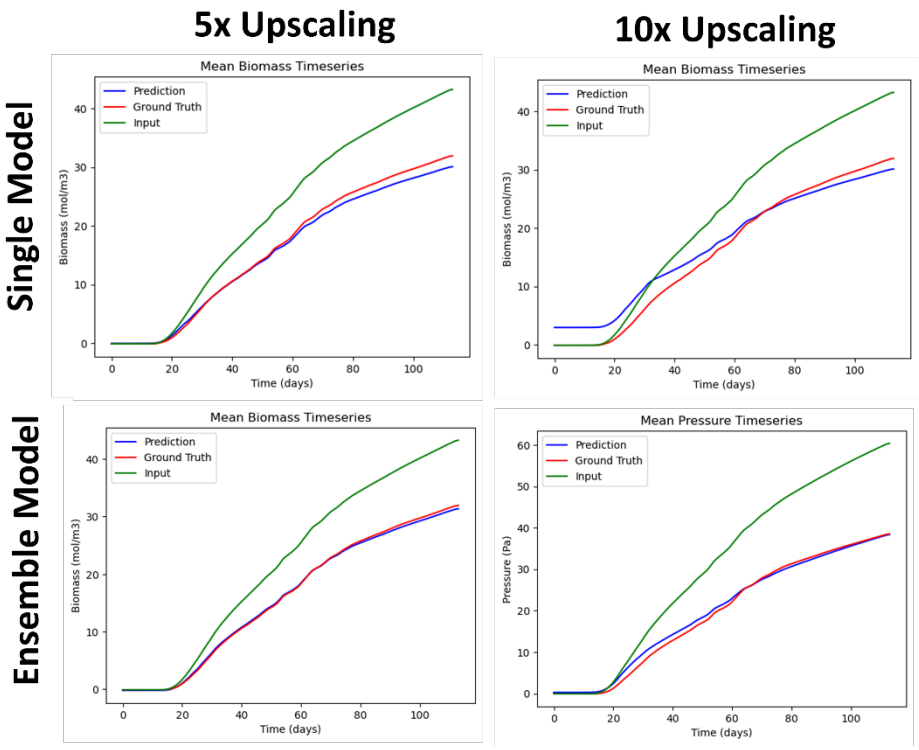

Figure 8. Results from biomass upscaling models. 10x upscaling represents an increase in the scale of the simulation domain by a factor of 10, and 5x upscaling represents an increase in the scale by a factor of 5. The ensemble model is a combination of single models with different parameters that results in improved generaltzation.

In addition to our biomass upscaling, we also trained models to upscale ED and Cr(VI). The 5x upscaling for both ED and Cr(VI) show good results, although the output from our ED model approximates the ground truth time series much more closely than that of Cr(VI). In addition, we can observe that the mean Cr(VI) spatial distribution for the final time step is not able to accurately replicate the ground truth. Although it is closer to the ground truth than the stretched-out input, and thus better than using no correction at all, the prediction from the DL model is not

able to capture the fine-grained spatial variations shown in the ground truth output. For ED, the prediction of the spatial distribution is more accurate, although it is still much noisier than the ground truth.

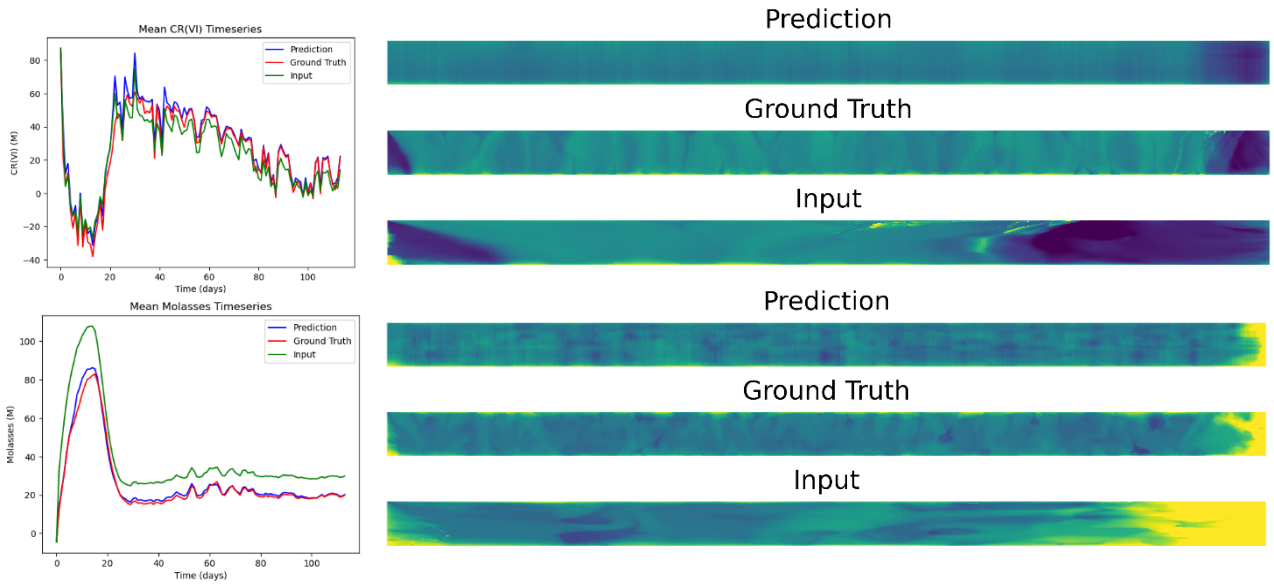

Figure 9. Multi Layer Perceptron (MLP) and Random Forest (RF) predictions vs true values for 30 day forecasting. MLP shows much better forecasting than RF at the 30, 60, and 90 day marks. RF is generally good at predicting low biomass values, but struggles to accurately forecast significant increases in biomass concentration.

Our work not only provides trained models for immediate upscaling, but also illustrates a proof of concept that can be used to design future upscaling models. First off, we show that training a model to upscale molasses can also allow for accurate predictions of biomass (**Fig. 10**). This surprising result indicates that our model can potentially be used to upscale features it wasn't trained on, and that highly correlated variables from RT simulations can potentially be used together to train a more robust model. Furthermore, we provide a general roadmap for this new method of upscaling in RT simulations, and a large amount of data and code that can be used to train better, or more general, DL-based upscaling models.

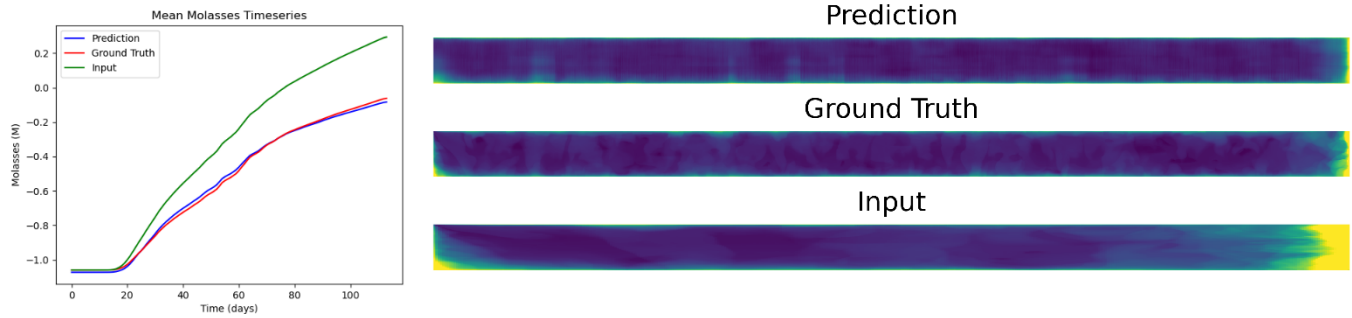

Figure 11. RF feature importances for 60 day forecast (left) and full time-series prediction from training on data with a different Keff (right).

4 Discussion