A machine-learning model to predict surface-wave amplification in sedimentary basins

Quentin Brissaud[1], Victor Tsai[2] and Daniel Bowden[3]
1: Caltech; 2: Brown University; 3: ETH Zurich

Caltech

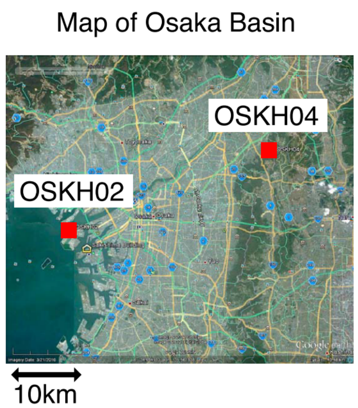
Motivations

- Surface waves (SW) are a known source of strong shaking in sedimentary basins

- There are three main approaches to study the basin response to SW
  1/ Analytical models – Inexpensive but oversimplified
  2/ Numerical simu. – Accurate but expensive
  3/ Real observations – limited good-quality dataset, many different phenomena involved

- Therefore, correlations between SW amplification and basin properties are poorly constrained

- Machine learning provides an inexpensive solution that accounts for nonlinear correlations between inputs and outputs



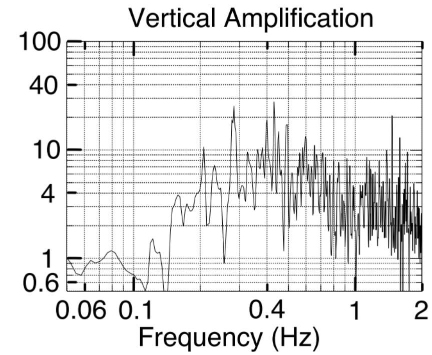


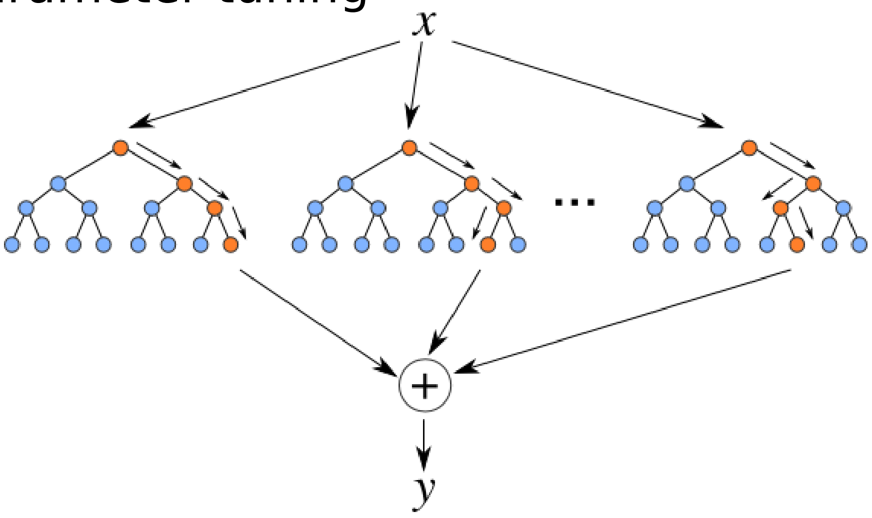
Figure: From *(Tsai et al, 2017)*
Vertical ampl. at station OSKH02 relative to OSKH04 from the Tohoku-Oki earthquake

Objectives

1. Can a machine-learning regression model accurately learn the nonlinear-correlations between the amplification spectra and the basin properties
2. How do machine-learning models perform to predict the Rayleigh-wave amplification using commonly-used proxies ?

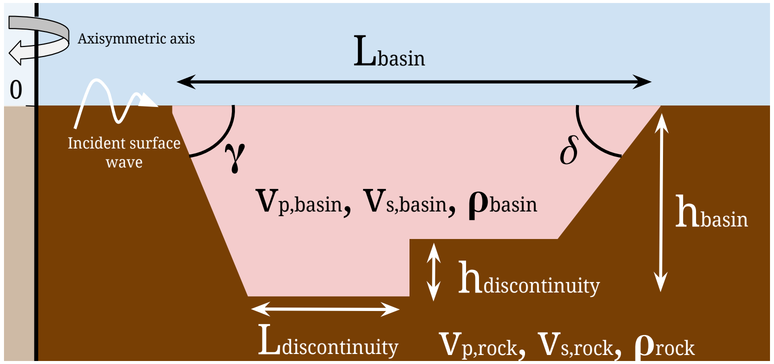
We train a **Random forest (RF)** to learn Rayleigh-wave amplification spectra for a given set of basin parameters**:**

- good generalization of training data
- provide insights about the feature importance
- limited hyper-parameter tuning



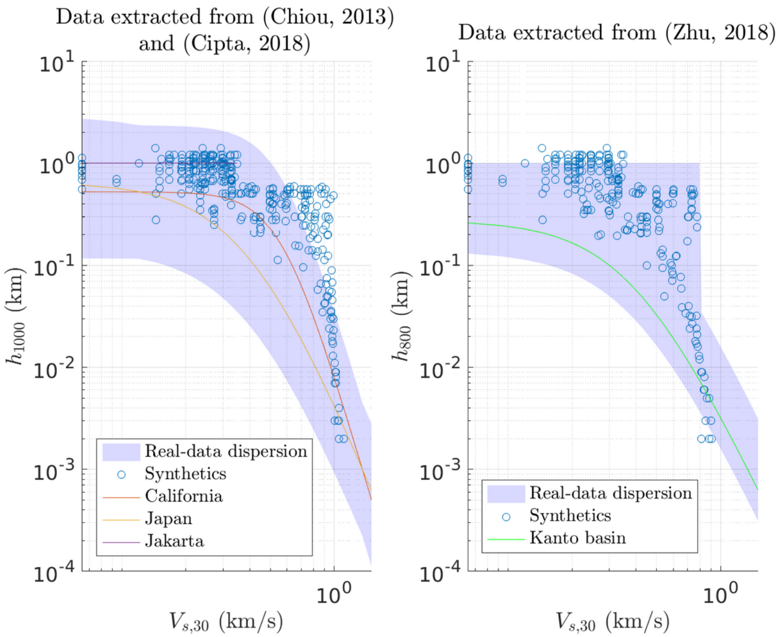
Dataset

- We generate a **training dataset from axi-symmetric numerical simulations**



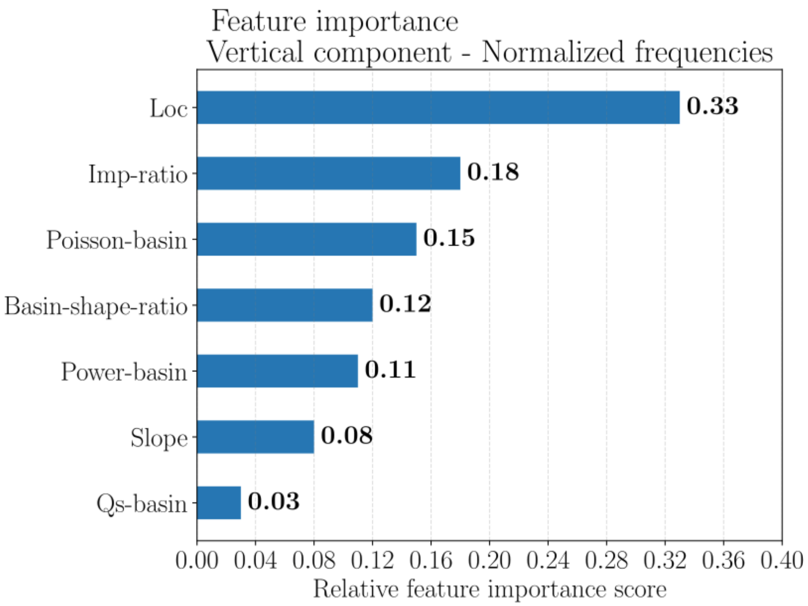
- We pick **basin parameters** in a reasonable range around empirical values

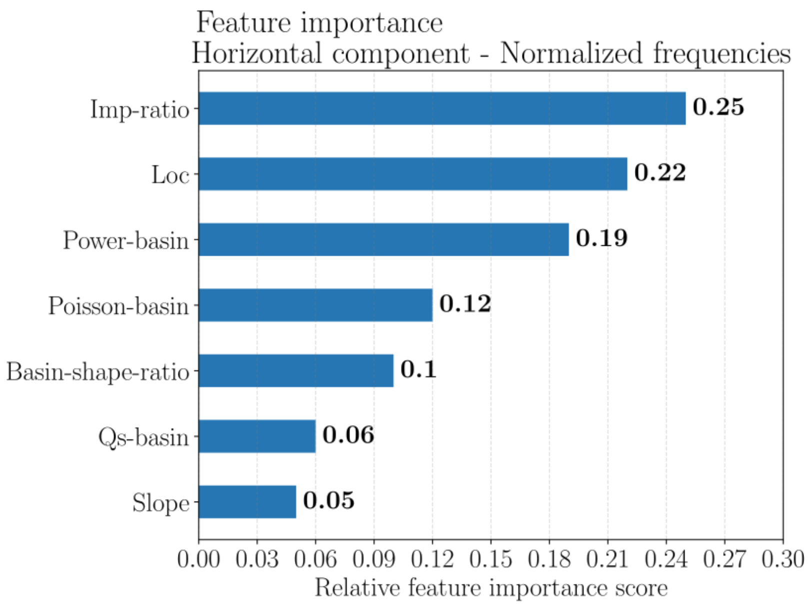
- Shear-velocity variations with depth follow **power laws**



Relative importance of basin parameters

- We train the RF to learn spectra over a normalized frequency range *(Cadet, 2012)* f=[0.1f0; 1.9f0], where f0 is the Rayleigh-wave dominant freq.
- **We only choose basin parameters that have a low Spearman's score** (S < 0.6)
- We obtain **0.93 (vertical) and 0.94 (horizontal) in R2 accuracy** using nondimensional basin parameters

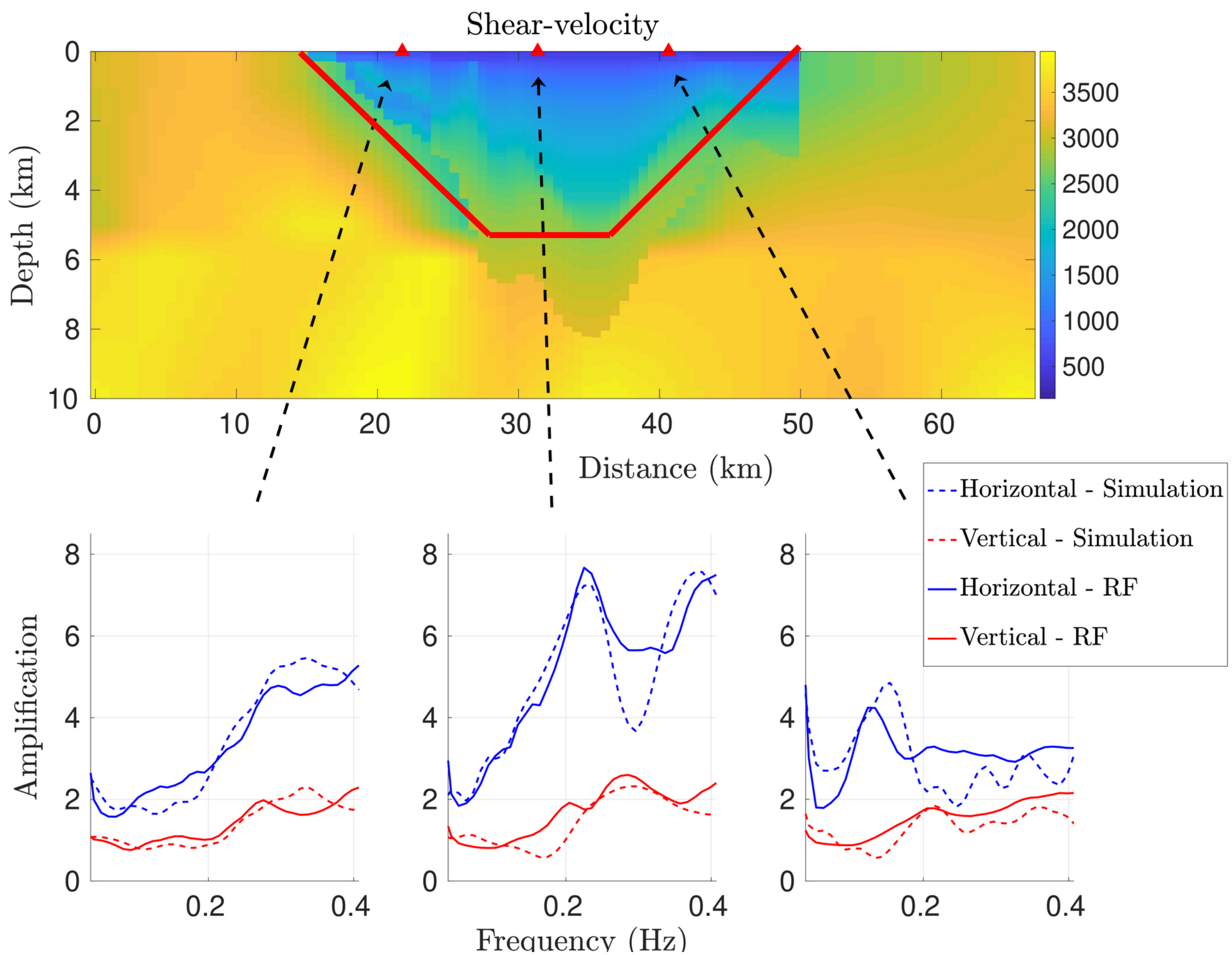




RF predictions in heterogeneous basins

- The main features of the amplification spectra in complex structures can be captured by RF over a limited frequency range

Basin-shape-ratio : 0.45
Imp-contrast : 0.33
Power-basin : 0.3
Slope : 0.5
Poisson-basin : 0.4
Qs-basin : 9999.0

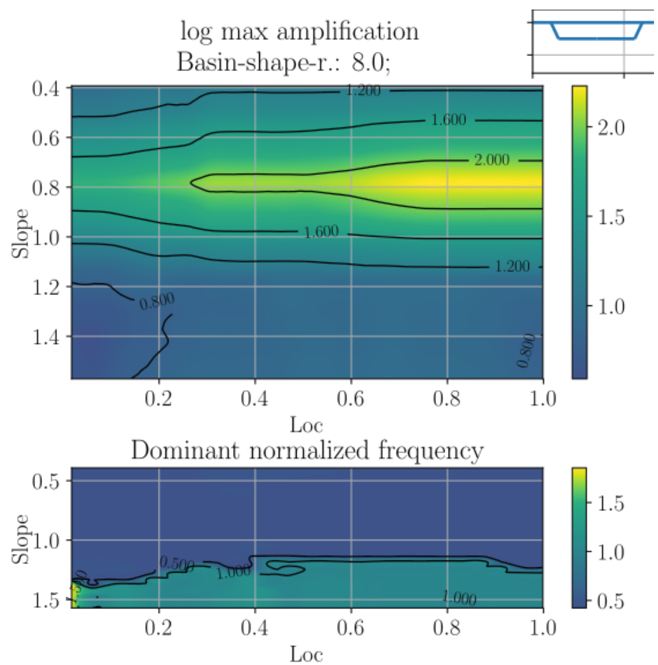


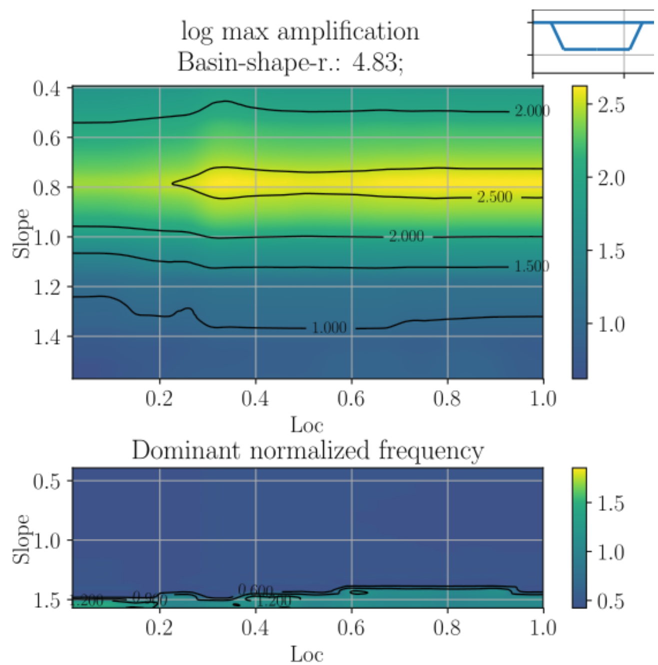
Basin-geometry impact on the maximum amp.

- The **highest amplifications occur for small basin-edge angles and deep basins** due to constructive interference between Rayleigh-wave modes close to the second basin edge.
- **Basins with low basin-shape ratios lead to larger amplification in the near-field** of the first basin edge
- Conclusions hold for the horizontal component

Vertical

Imp-ratio : 6.46
Power-basin : 0.0
Poisson-basin : 0.24
Qs-basin : 9999.0



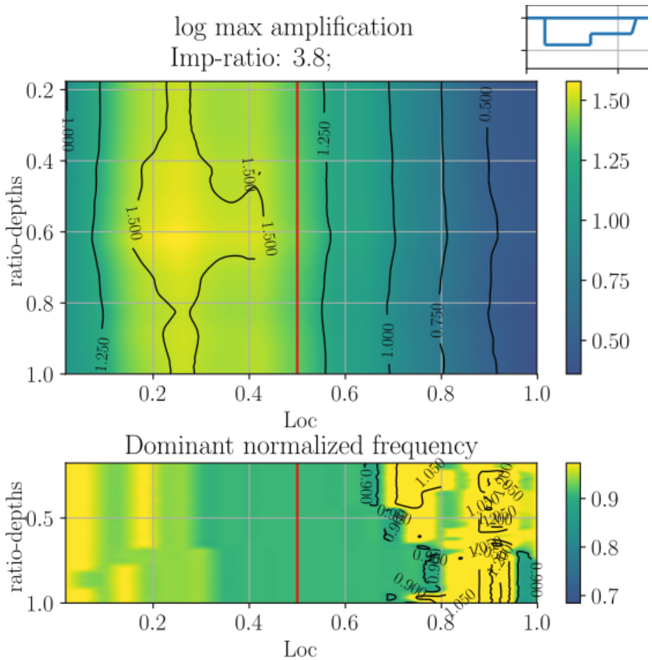


Basin-geometry impact on the maximum amp.

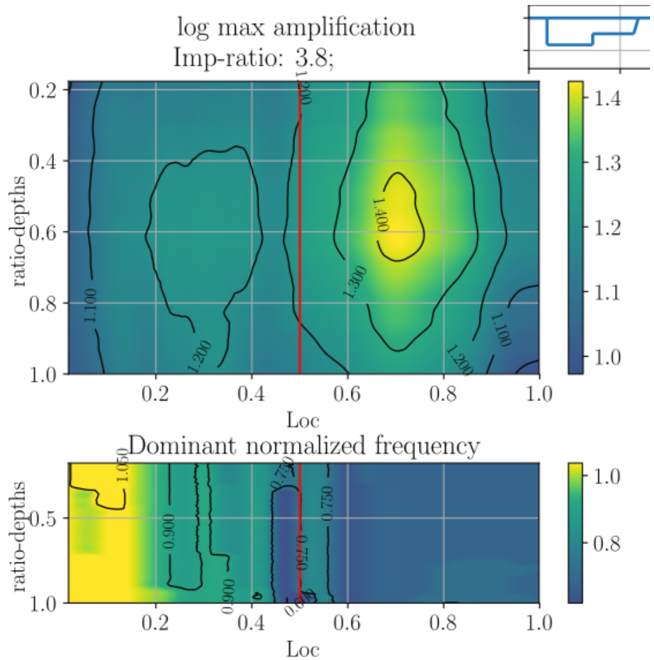
- **The horizontal amplification can be significantly increased in the shallower part of the basin** owing to diffracted/converted waves at the basin discontinuity

Basin-shape-r. : 4.83
Imp-ratio : 3.8
Power-basin : 0.0
Slope : 1.57
Slope-2 : 1.18
Poisson-basin : 0.24
Qs-basin : 9999.0

Vertical



Horizontal



Performance of commonly-used proxies

- **Most proxies used so far are not generally accessible** at real sites. Instead, other proxies such as vs30 or h800, are used to infer the ground motion
- **By training RF models with different sets of parameters we can compare the performance of commonly-used proxies** to SW amplification

Relative location in the basin

Frequency at max. H/V ratio

Imp. ratio at depth h1500

Depth where Vs = 1.5 km/s

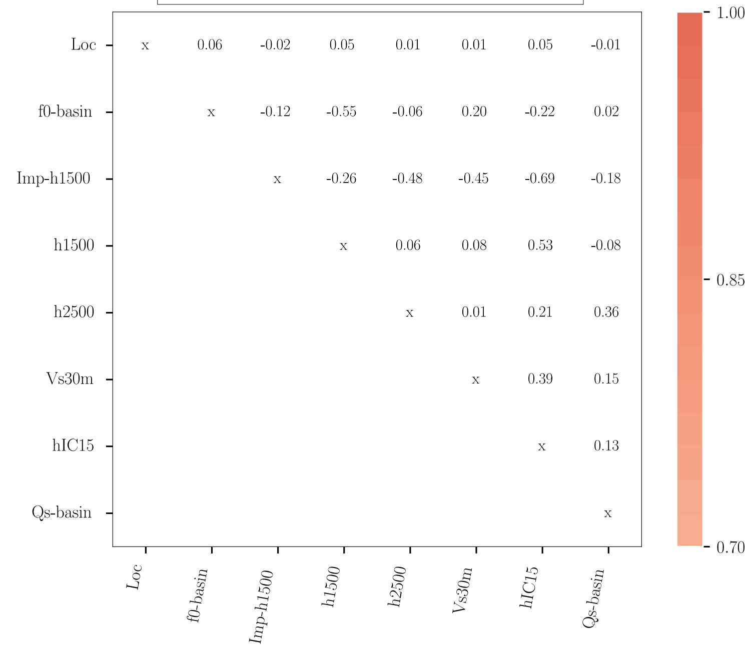
Depth where Vs = 2.5 km/s

Average Vs at depth h = 30 m

Depth where the Imp. Ratio > 1.5

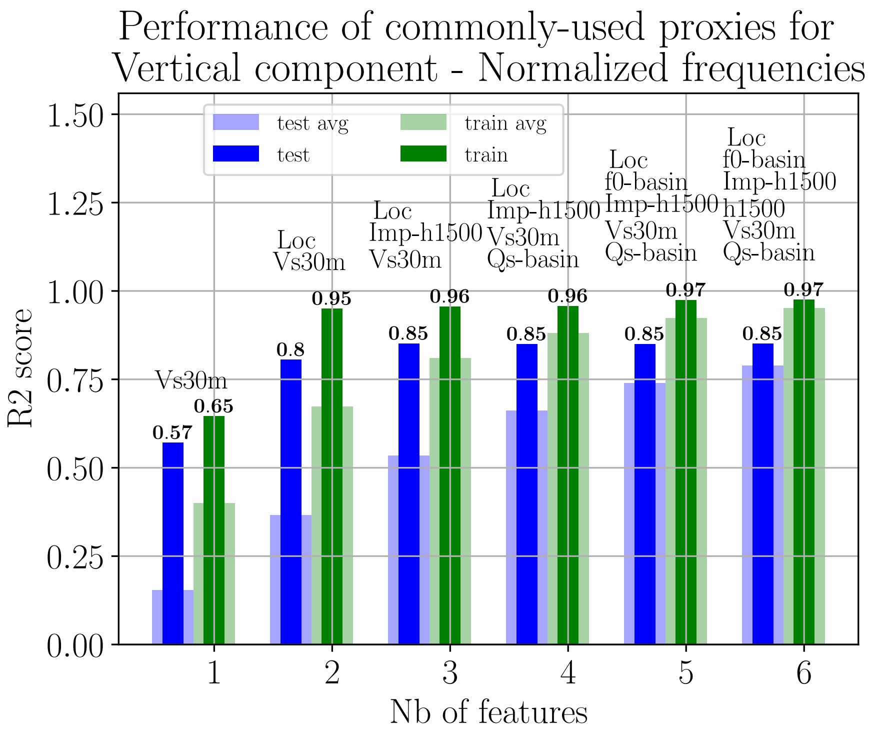
Shear quality factor

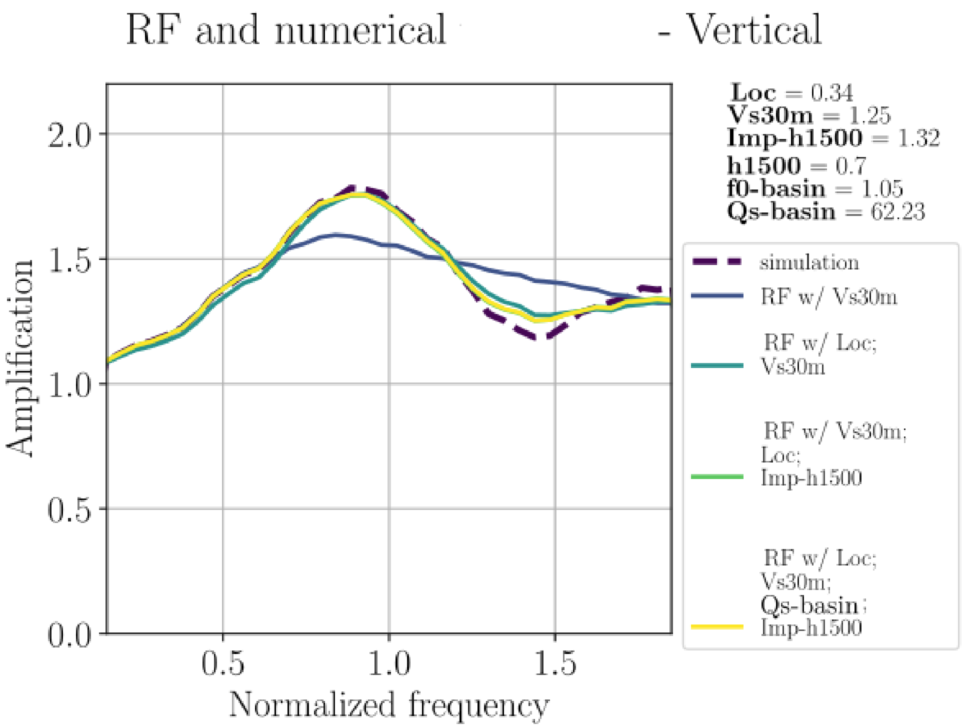
Spearman's matrix



Performance of commonly-used proxies

- Relative location significantly increases the R2 accuracy
- Performances reach a plateau when the number of features exceeds 3
- Note that performances are similar for the horizontal component





Caltech

Conclusions

- **A RF trained on synthetics provides an accurate regression model** for simple basin structures to predict SW amplification. RF importance and performance metrics support previously-published results
- **We observe that amplification maxima in a simple basin are strongly correlated to the basin-shape ratio and the basin-edge slope**. By neglecting SW resonance and geometrical effects we might underpredict wave amplification.
- **Normalizing spectra using a non-dimensional frequency greatly increases the accuracy** when a strong basin discontinuity is present.
- Relative location within the basin can improve the performance of regression models
- A RF trained over synthetics could improve the basin response to shallow crustal events in ground-motion models

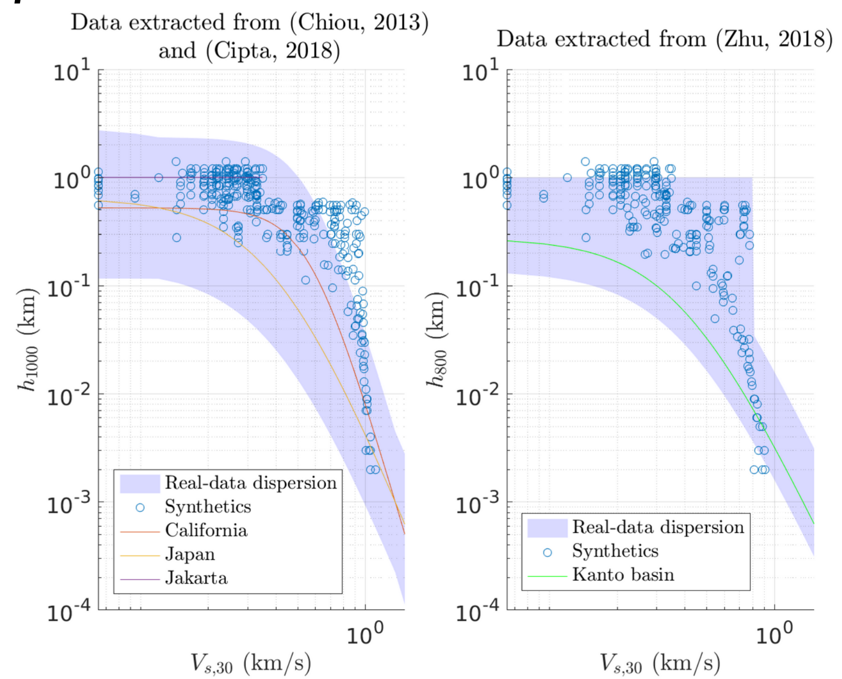
References

- Chiou, B. S. J., & Youngs, R. R. (2014). Update of the Chiou and Youngs NGA model for the average horizontal component of peak ground motion and response spectra. Earthquake Spectra, 30(3), 1117-1153.
- Cipta, A., Cummins, P., Irsyam, M., & Hidayati, S. (2018). Basin resonance and seismic hazard in Jakarta, Indonesia. Geosciences, 8(4), 128.
- Zhu, C., Chávez-García, F. J., Thambiratnam, D., & Gallage, C. (2018). Quantifying the edge-induced seismic aggravation in shallow basins relative to the 1D SH modelling. Soil Dynamics and Earthquake Engineering, 115, 402-412.
- Cadet, H., Bard, P. Y., Duval, A. M., & Bertrand, E. (2012). Site effect assessment using KiK-net data: part 2—site amplification prediction equation based on f 0 and Vsz. Bulletin of Earthquake Engineering, 10(2), 451-489.
- Tsai, V. C., Bowden, D. C., & Kanamori, H. (2017). Explaining extreme ground motion in Osaka basin during the 2011 Tohoku earthquake. Geophysical Research Letters, 44(14), 7239-7244.

Appendix

Dataset

- **Synthetics basin parameters** are chosen in a reasonable range around values collected by (Chiou and Youngs, 2013), (Cipta, 2018) and (Zhu, 2018)



- **Attenuation** is computed from empirical model by (Graves and Pitarka 2010)
- **Vp-Vs-Density** relationships used are from (Brocher, 2005)

RF Training to investigate proxy performance

- The performance of proxies with dimensionalized values (e.g. h1000, vs30) is **frequency-dependent**. Each discontinuity with depth impacts the RW spectrum over a specific freq. range (Brissaud and Tsai, 2019). If a proxy can capture the discontinuity properties (e.g. depth, imp. Ratio) then the proxy will be efficient at predicting the amp. spectrum features.

- **There is a strong correlation between the frequency at maximum amp., the basin depth and the shear velocity above the basin main discontinuity**. Using a non-dimensional frequency based on these correlations improves the regression model performance as pointed out by (Cadet, 2012) and facilitates comparisons b/w datasets.

- Various input features can be selected to train the RF. However, **strongly-correlated features** will: **1/** make the feature importance metric bias and **2/** Slow down the training process

- Performance is traditionally measured either as a variance reduction from a given initial model or a variance residual b/w the observed data and predicted data. Here **we use the R2 metric, that corresponds to the % of variance explained by a model. R2 is independent of the amplification absolute value**.

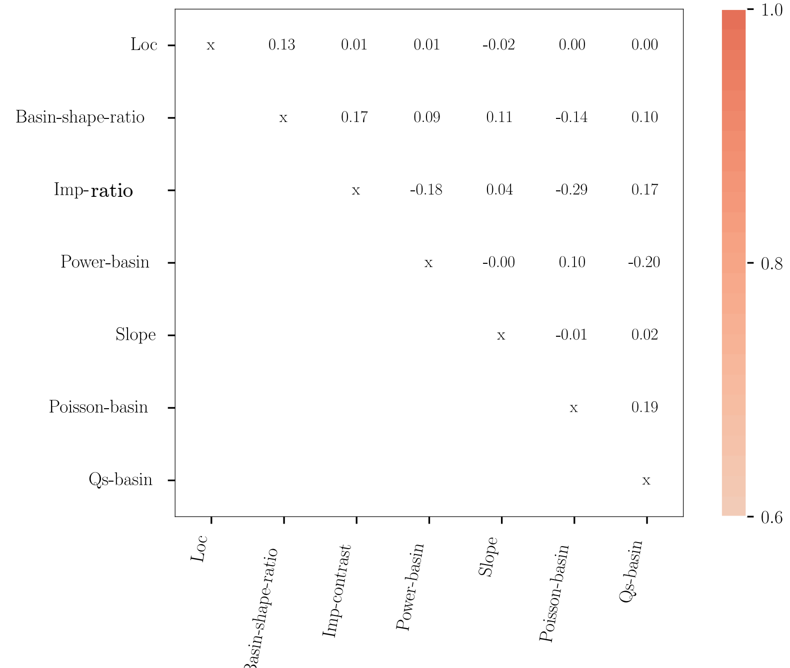
Appendix: Correlations for best proxies

- To study the significant correlations between the basin parameters and the SW amplification, we choose an **input feature set with the highest R2 score and lowest Spearman's coefficient**.

- Best set of non-correlated input features:

Basin-shape ratio
Imp. ratio at basin discontinuity
Power from shear-wave veloc. power-law model
Basin-edge slope
Poisson's ratio in the basin

Spearman's matrix



- Note that these features are only relevant for theoretical purposes since they are usually not well-defined for real basins

Appendix: Correlations for common proxies

- We removed **strongly-correlated features** defined by a Spearman's coefficient > 0.7 (Jusoh, 2017)

- We consider the following input features:

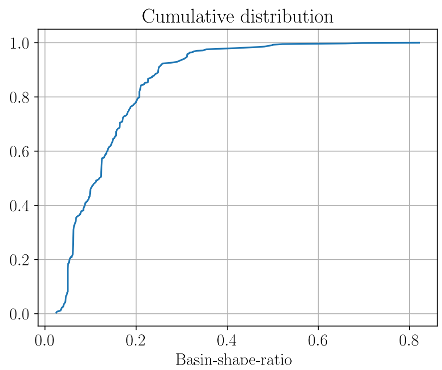
Relative location in the basin ◄— Loc
Frequency at max. V/H ratio ◄— f0-basin
Imp. ratio at depth h1500 ◄— Imp-h1500
Depth where Vs = 1.5 km/s ◄— h1500
Depth where Vs = 2.5 km/s ◄— h2500
Vs at depth h = 30 m ◄— Vs30m
Depth where the Imp. Ratio > 1.5 ◄— hIC15
Shear quality factor ◄— Qs-basin

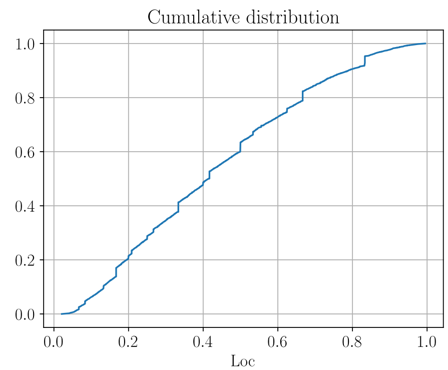
	Loc	f0-basin	Imp-h1500	h1500	h2500	Vs30m	hIC15	Qs-basin
Loc	x	0.06	-0.02	0.05	0.01	0.01	0.05	-0.01
f0-basin		x	-0.12	-0.55	-0.06	0.20	-0.22	0.02
Imp-h1500			x	-0.26	-0.48	-0.45	-0.69	-0.18
h1500				x	0.06	0.08	0.53	-0.08
h2500					x	0.01	0.21	0.36
Vs30m						x	0.39	0.15
hIC15							x	0.13
Qs-basin								x

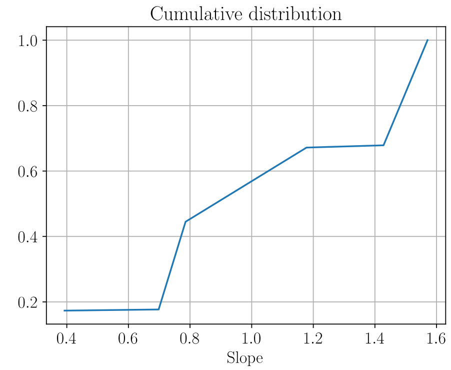
Color scale: 0.70 – 0.85 – 1.00

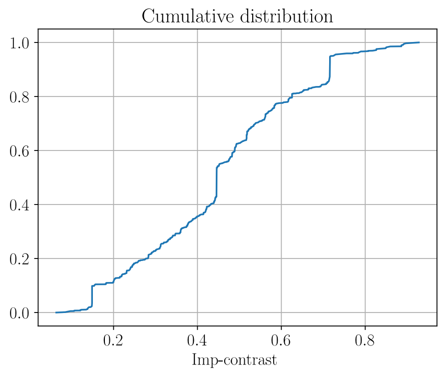
- H1000 and Vs10 have strong correlations with h1500 and Vs30 and are therefore not included

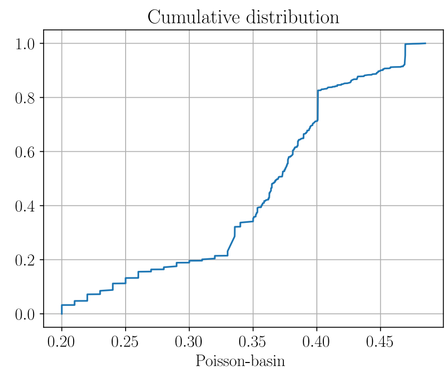
Appendix: Input cumulative distribution

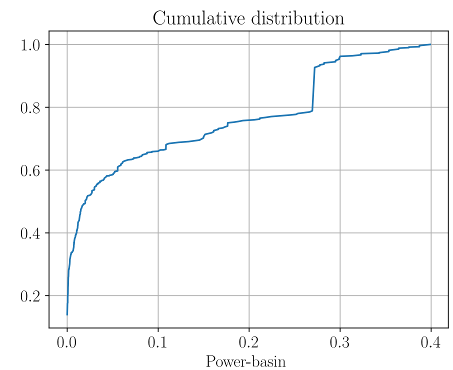


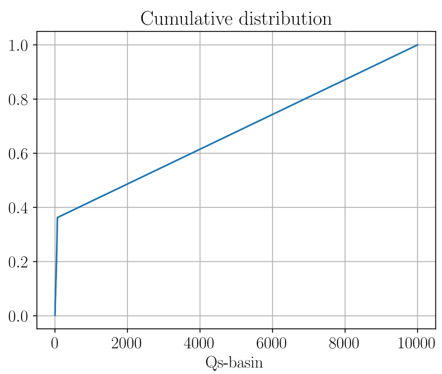




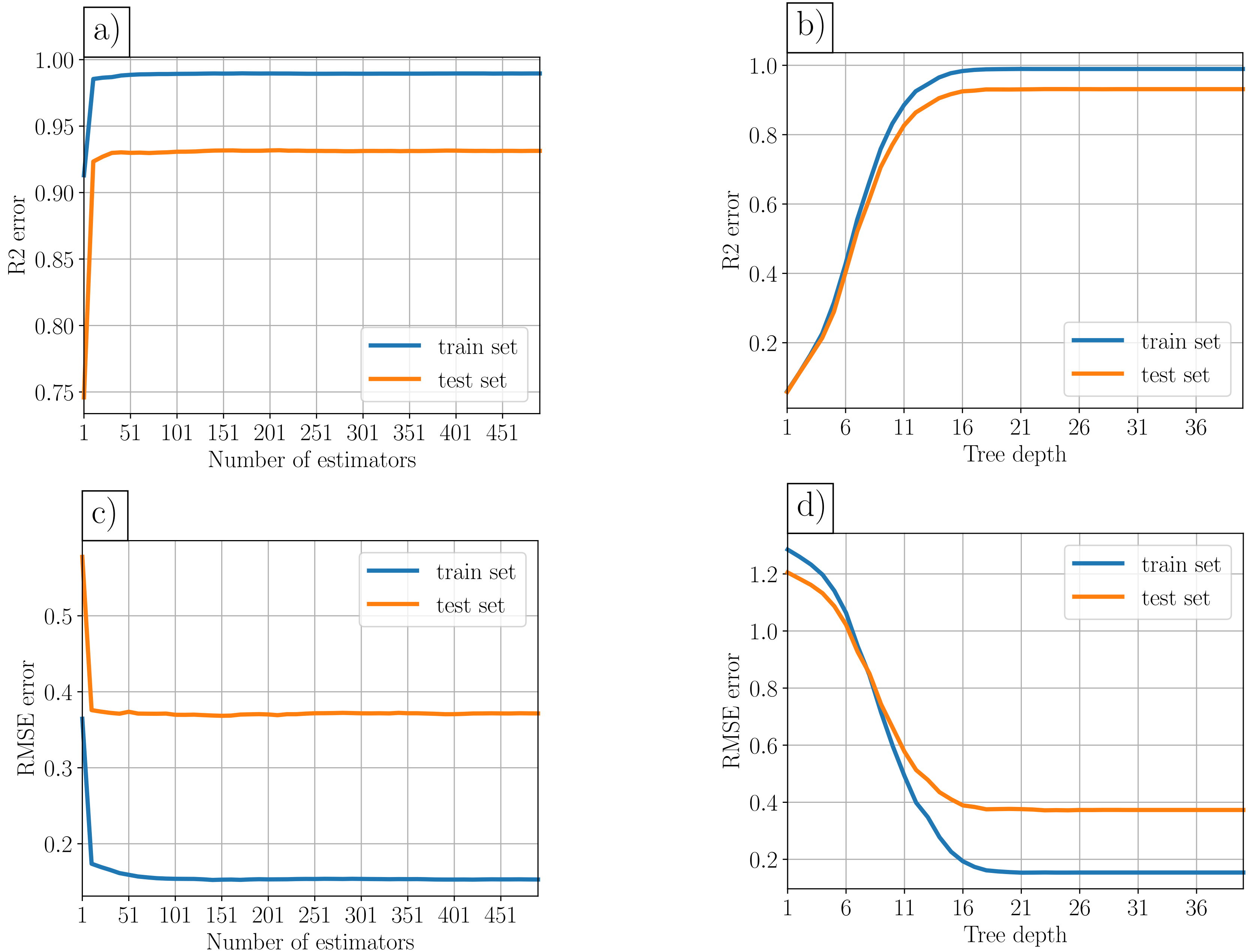








Appendix: Convergence tests



Notes about the training data and the perf. comp

- The spectra are either computed over a “non-dimensional” fre range or a real freq. range. The **non-dimensional freq. rang is [0.3*fh; 1.7*fh] Hz**, with fh the non-dimensional freq. suc that fh = Vsh/(2.25*h) where Vsh is the shear velocity above the main discontinuity and h the depth at which the basin is t deepest. The **real freq. Range is [0.01; 2.5] Hz**

- The performance comparison plots (bar plots) show the best c features, **for a given number of input features**. The perfor training and testing datasets are shown in order to see when t overfits (when the perf. of traning dataset is significantly larg dataset).

- The bar plots also show the average performance computed fr all the proxy combinations. It indicates whether or not any combination of features would have a good performance

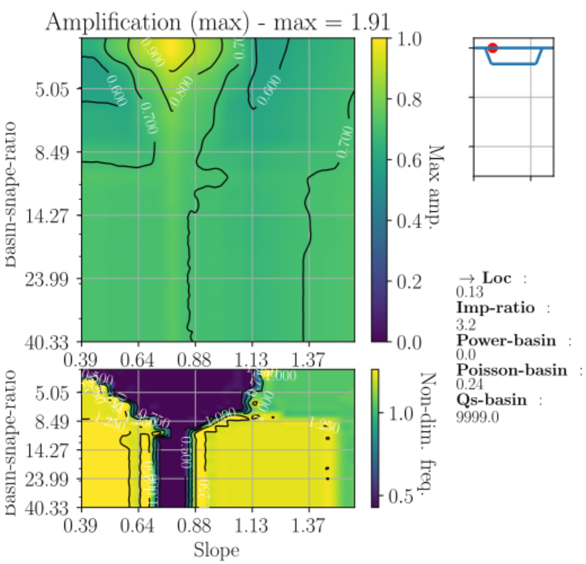
Notes about the feature importance

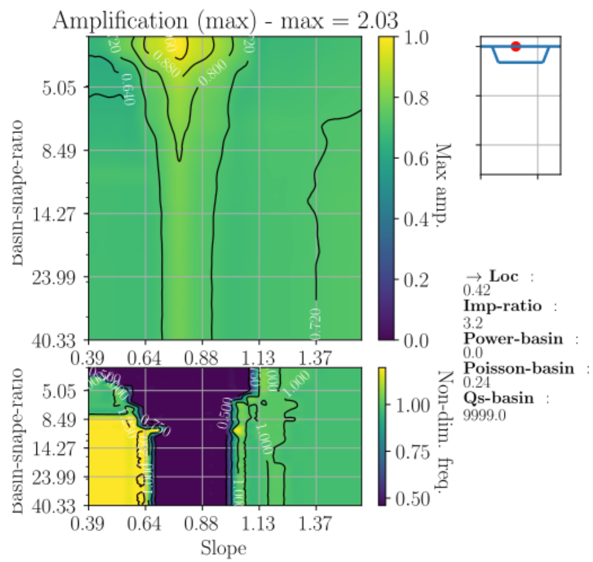
- The feature-importance score is an inherent metric of RF. Importance measures, for a set of independent input variables the effect that a given input has on the output relatively to oth inputs. If an input is important, changing its value alters the output values more significantly than changing other less important inputs.

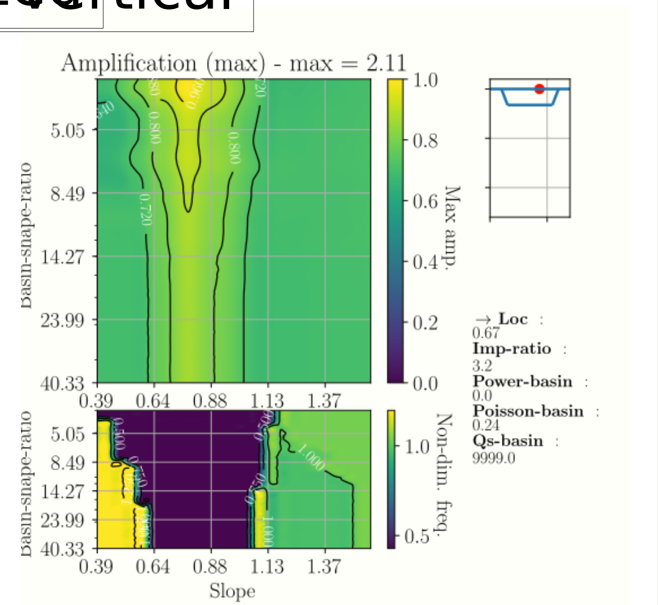
How does the basin geometry impacts the maxim

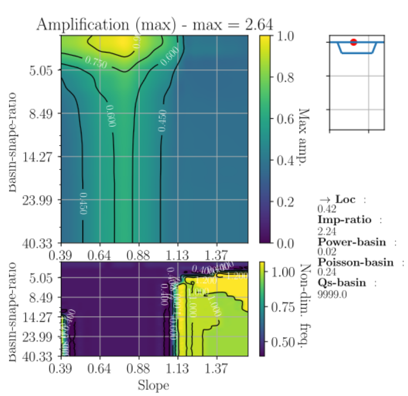
- There is a combination of basin-shape ratio and basin slope
  global maximum at lower frequencies for a given Imp. ratio

Amplification is normalized Vertical









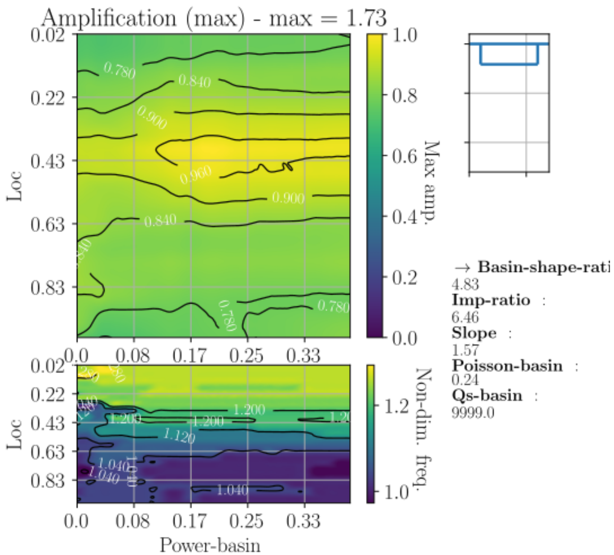
- For a given Poisson's ratio, the maximum amp
  against basin-shape ratio and basin slope is in
  of Imp. ratio.
- The max. amp. does not monotonously increas
  increasing Imp. Ratio because there is a
  trade-off between freq. at max. amp. and
  basin-shape ratio

How do vertical and horizontal amplification diffe

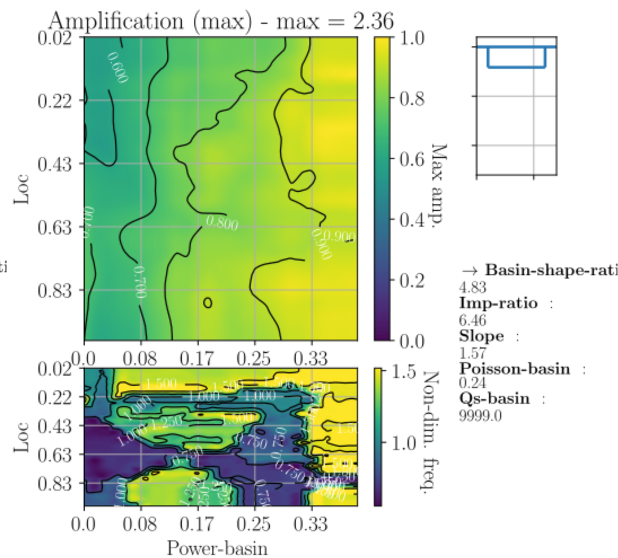
Amplification is normalized

- For given basin-shape ratio and Imp. ratios, the max. amplification tends to be homogeneous for a given loc. for the vertical comp. and for a given power for the horizontal comp.

Vertical



Horizontal



- As pointed out earlier: "The generally stronger fundamental-to-first horizontal trans. coef. might explain the larger importance of the power exponent."