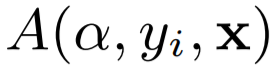
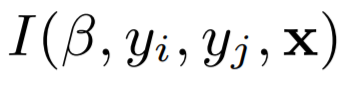
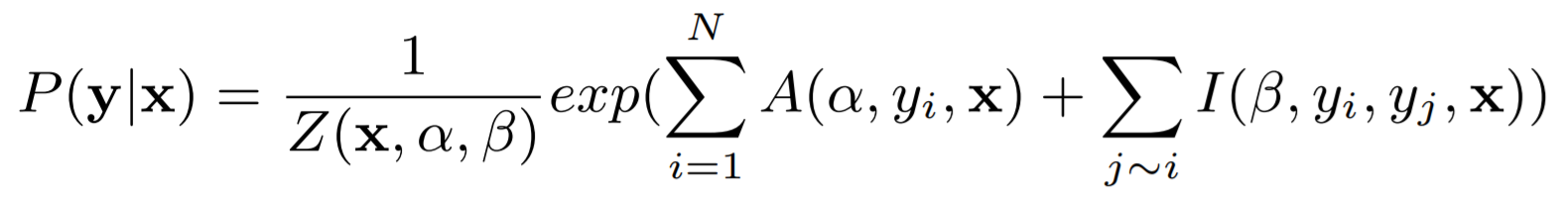
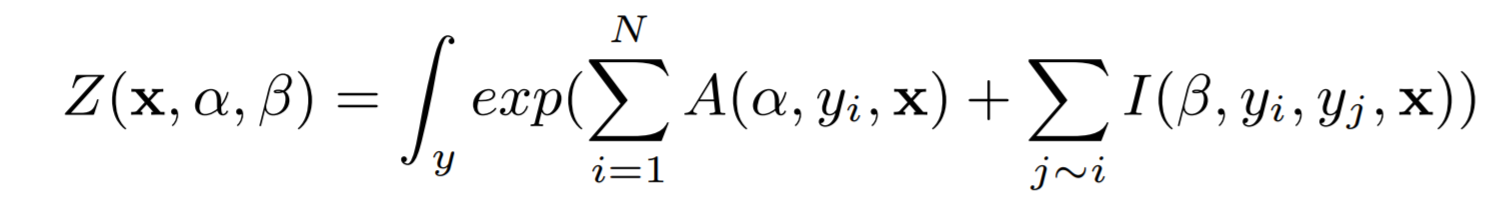
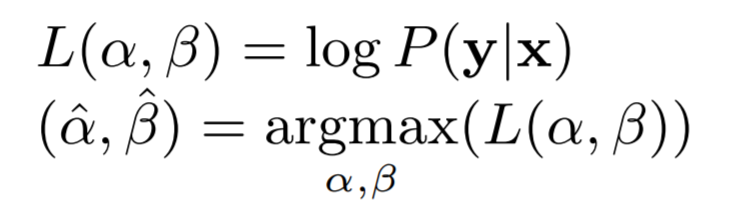
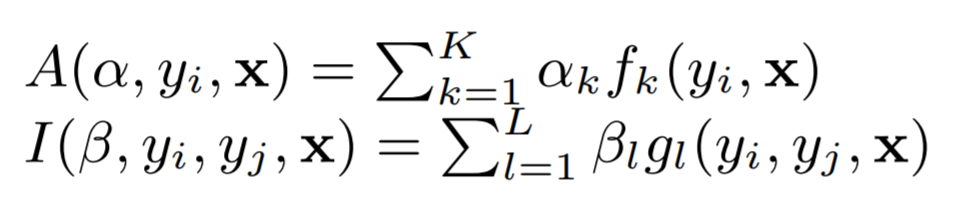
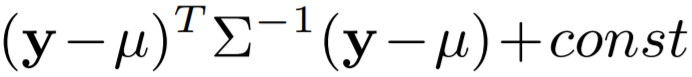
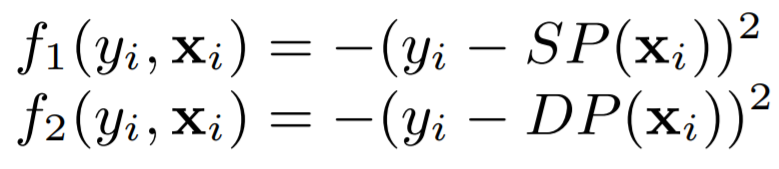
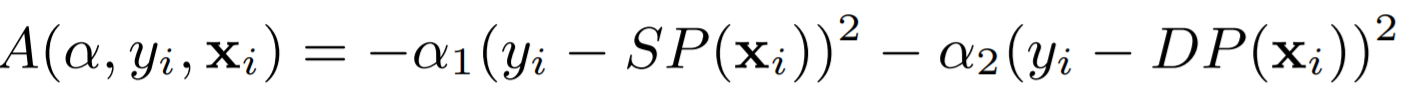
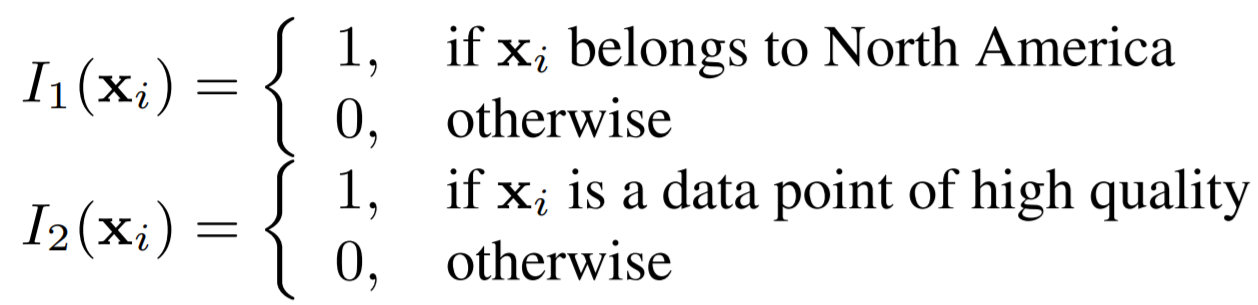
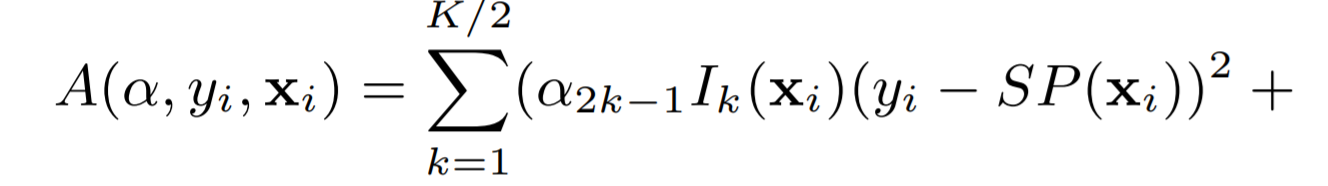
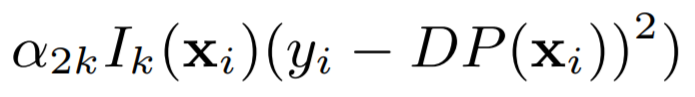
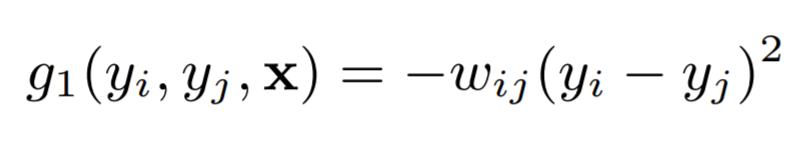
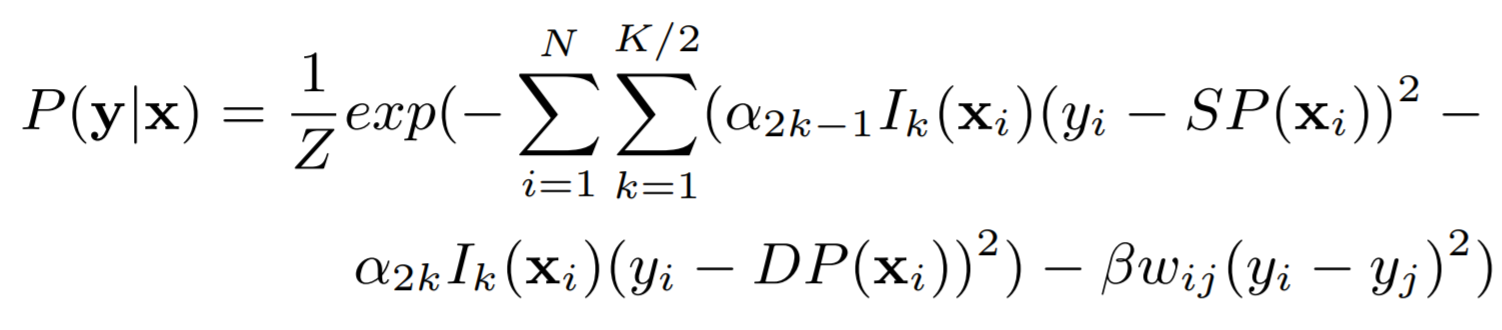
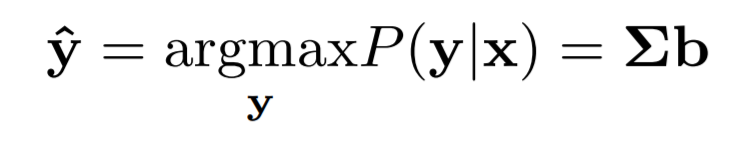
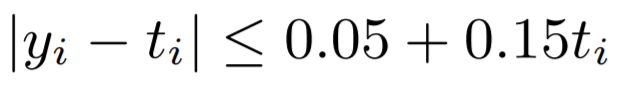
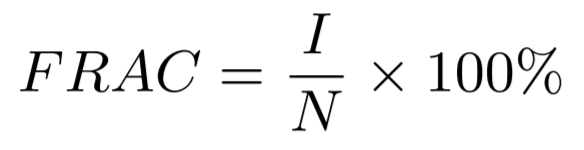
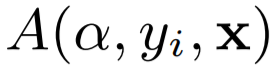
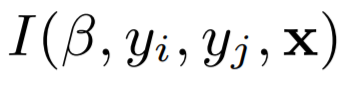
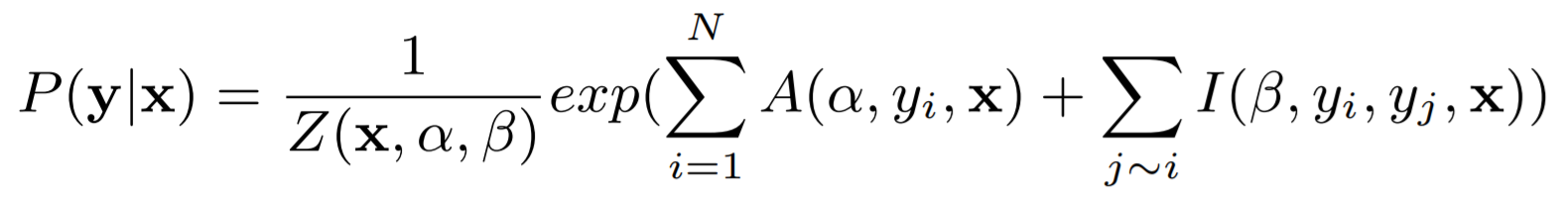
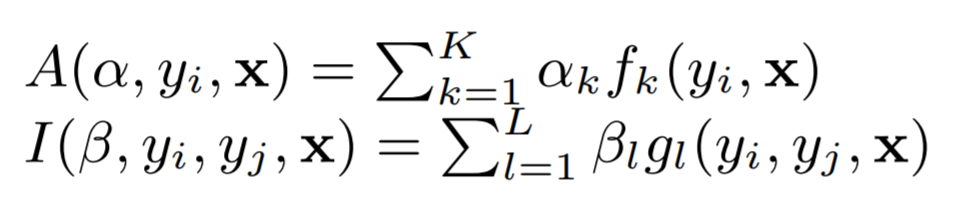
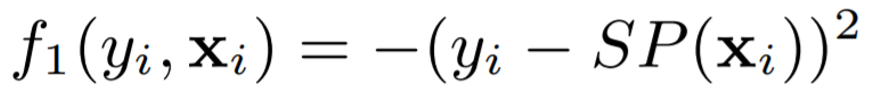
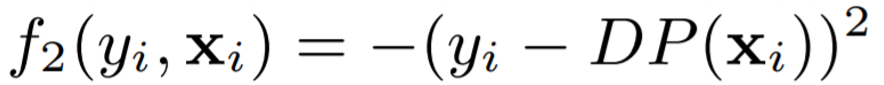
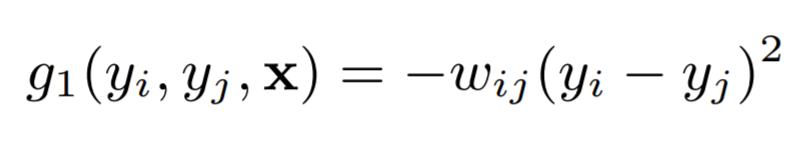
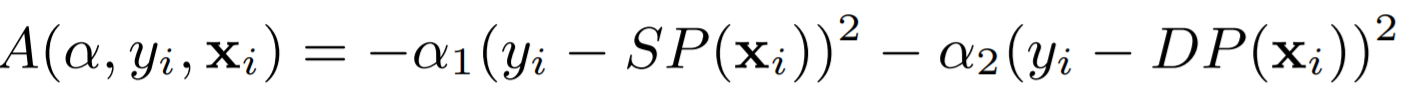
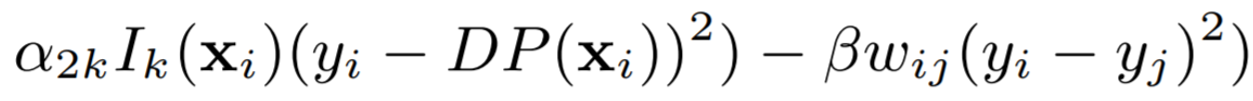
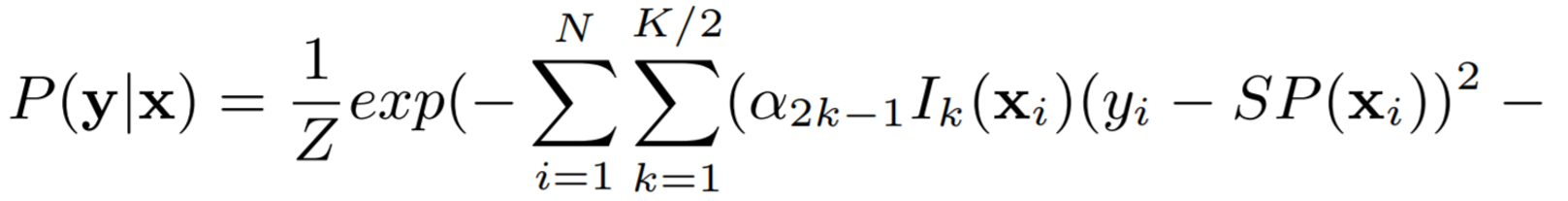
**Paper Review: “Continuous CRFs for Regression in Remote Sensing”**

* CRFs are widely used for predicting output variables that have some internal structure
* Most CRF research is done on structured classification where outputs are discrete
* Paper proposes CRF probabilistic model for structured regression that uses multiple non-structured predictors as features
  + Construct features as squared prediction errors and show that this results in a Gaussian predictor
* Learning becomes a convex optimisation problem leading to a global solution for a set of parameters
* Inference can be conveniently conducted through matrix computation
* Experimental results on remote sensing problem of estimating Aerosol Optical Depth (AOD) provide strong evidence that proposed CRF model successfully exploits inherent spatio-temporal properties of AOD data
  + Experiments reveal that CRF is more accurate than the baseline neural network and domain-based predictors
* Traditional supervised model that uses only information contained in a single sample (which is noisy), ‘xi’, to predict output, ‘yi’, might predict the value of ‘yi’ to be quite different to that of ‘y­­i-1’ and ‘yi+1’ because it treats them individually
* Structured predictor uses dependencies among outputs to take into account that ‘y­I’ is more likely to be closer to ‘y­i-1’ and ‘yi+1’, thus improving final predictions
* Usually have prior knowledge about relationships among outputs ‘y’
  + Usually application-specific where dependencies defined in advance, either by domain knowledge or by assumptions
* Relationships among outputs can be represented by graphical models
  + For spatial-temporal data, includes MRFs and CRFs
* Originally CRF designed for classification of sequential data
  + Recent applications in computer vision and computational biology
* Paper builds on continuous CRF model to develop a solution for spatio-temporal data
* In contrast to traditional domain-drive approaches to AOD prediction algorithms, paper proposes a completely data-driven approach, which consists of training non-linear regression model to predict AOD using satellite obs as inputs, while targets are obtained from distributed ground-based sites over the world
* Property of statistical prediction is that high accuracy guaranteed only for the conditions similar to those of the ground-based sites
  + However, some regions are underrepresented due to non-uniform distribution of ground-based sites over the world
* Goal is to combine the two approaches so that in some regions where we know deterministic algorithm performs better, we rely more on deterministic model than on a statistical method, and vice versa
* Aerosol data is characterized by strong spatial and temporal dependencies that CRF are able to exploit by defining interactions among outputs using feature functions
* Use of features to define CRF models allows us also to include arbitrary properties of input-output pairs into the compatibility measure
* Paper proposes CRF probabilistic model for structured regression that uses multiple non-structured predictors as its features
  + Features as squared prediction errors of deterministic and statistical models and shows that results in multivariate Gaussian conditional distribution P(y|x), where learning becomes convex optimisation problem leading to a global solution
* CRF provides probabilistic framework for exploiting complex dependence structure among outputs by directly modelling conditional distribution P(y|x)
* Association potential function  associates the output ‘yi’ with input vectors ‘x’ by real-valued function with ‘α’ being a K-dimensional set of params
  + Larger value of ‘A’ results in more ‘yi’ being related to ‘x’
  + Usually ‘A’ is a combination of functions
* In general ‘A’ takes all input data ‘x’ to predict a single output ‘yi’, meaning it doesn’t impose any independency relations among inputs ‘xi’ and can also use as many association functions as we feel necessary to model input-output relations in data
* To model interactions among outputs, real-valued function called the interaction potential  is used, where ‘B’ is L-dimensional set of parameters
  + Represents relationship between two outputs and in general can depend on the input ‘x’
  + E.g. in AOD prediction problem, can be modelled as correlation in time and space between neighbouring outputs
  + The larger the value, the more related the outputs are
* CRF models conditional distribution ‘P(y|x)’, y=(y1, …, yN) according to associated graphical structure
* 
* Where Z is normalisation function defined as
* 
* Learning task is to chose values of parameters ‘α’ and ‘B’ to maximise the conditional log-likelihood of set of training examples
* 
* Can be achieved via standard optimisation algorithms, e.g. gradient descent
  + Regularise ‘L(α, B)’ by adding ‘α2/2’ and ‘B2/2’ terms to argmax formula that prevents params from becoming too large
* Models w/ real-value targets pose different challenges to discrete case
  + Most important difference is that the normalisation function ‘Z’ is integral, not sum
  + Discrete valued models always feasible as ‘Z’ is finite number defined as sum of finitely many possible values of ‘y’
  + On the contrary, to have a feasible model with real valued outputs, Z must be integrable
* In CRF applications, ‘A’ and ‘I’ association and interaction potentials could be defined as linear combinations of a set of fixed features in terms of ‘α’ and ‘B’
  + 
* Use of features to define model is convenient because it allows us to include arbitrary properties of input-output pairs into compatibility measure; this way, any potentially relevant feature could be included to the model because parameter estimation automatically determines actual relevance by feature weighting
* In general, to evaluate ‘P(y|x)’ needed during training and inference, one would need to use time consuming sampling methods such as MCMC-based algorithms
  + However, if ‘A’ and ‘I’ are defined as quadratic functions in terms of ‘y’, then sum ‘A’+’I’ can transform to 
  + Expression corresponds to multivariate Gaussian distribution with mean ‘μ’ and covariance 
* If this is convex, learning and inference are convenient
* Learning becomes a convex optimisation problem leading to a global solution for ‘α’ and ‘B’ and inference can be conveniently conducted through matrix computation
* Given new observation ‘x’ for inference, output ‘y’ calculated as conditional expectation ‘E(y|x)’
* By exploiting sparsity inherent to spatio-temporal, inference can be performed in time linear with number of spatial-temporal observations
* Need ‘P(y|x)’ to be feasible conditional distribution
  + Condition for ‘P(y|x)’ to be multivariate Gaussian is that normalization function ‘Z’ is finite (i.e. if covariance matrix  is positive semi-definite), hence when learning parameters we have to imply constraint that covariance matrix is semi-definite
* Given dataset that consists of satellite observations and ground-based AOD measurements, statistical prediction model (SD) can be trained to use satellite observations as attributes and predict the labels which are ground-based AODs
  + Deterministic AOD prediction models (DP) based on solid physical principles and tuned by domain scientists
* To model association potential (i.e. dependency between preds + target AOD), introduce 2 feature functions
  + 
  + Where for a given observation ‘xi’, ‘SP(xi)’ and ‘DP(xi)’ are outputs of statistical and deterministic models, respectively
* These follow basic principle for association potentials (values larger for more accurate predictions)
* Learned params ‘α’ of linear combination of these features
  + 
  + Provide some insight on how much to trust the SD and DP prediction algorithms
* E.g. large ‘α1’ places large penalty on mistakes of SD model and is indicator of this predictor
* Indicator functions can be:
* 
* Association potential now becomes
*  
* By introducing indicator functions we essentially partition whole dataset into smaller subsets
  + Learned ‘α’ represents our belief in SD and DP in different subsets, corresponding to different prediction conditions
* To model interaction potential, introduces a feature function:
* 
* ‘wij’ positive number representing a measure of spatio-temporal proximity between data points ‘i’ and ‘j’
  + Closer points are given larger weight
* Resulting CRF model is thus:
* 
* Above equation can now be represented as a multivariate Gaussian distribution
* To ensure that the model is feasible, when putting the above equation in Gaussian multivariate distribution form, the covariance matrix  has be to positive semi-definite, which means that all params have to be >0
* Learning is thus a constrained optimisation problem because we need to guarantee that ‘ak’>0 and ‘B’>0
  + Gradient ascent cannot be directly applied to a constrained optimisation problem
* To get around this, maximise log-likelihood w.r.t. ‘log ak’ and ‘log B’ instead of ‘ak’ and ‘B’
  + As a result, new optimisation problem becomes unconstrained
* In inference, since the model is Gaussian, prediction will be expected value, which is equal to the mean ‘μ’ of the distribution
  + 
* Paper considers data from MODerate resolution Imaging Spectrometer (MODIS), an instrument aboard NASA’s Terr and Aqua satellites
  + Instruments mounted on Terra observe Earth during morning, whereas those on Aqua observe Earth during afternoon
* Study uses only Terra satellite data
  + Ground-based data obtained from AErosal RObotic NETwork (AERONET), which is a global remote sensing network of radiometers that measure AOD several times per hour from specific geographic locations
* MODIS has high spatial resolution (pixel small as 250x250 m2 and achieves global coverage daily)
* However, AERONET sites, situated at fixed geographical locations, achieve data at intervals of 15 mins on avg
  + Gives rise to need for both spatial and temporal data fusion
* Fusion method involves aggregating MODIS pixels into blocks of 50x50 km2 and spatially collocating them with AEORNET site
* MODIS observations are said to be temporally collocated with corresponding AERONET AOD predictions if there is a valid AOD prediction within 30 mins of satellite overpass
* To avoid potential problems with outliers in ground truth data, AERONET level 2.0 observations were considered since they were cloud screened and manually verified
* Paper extracted satellite-based attributes from MODIS observations (collocated with AERONET level 2.0 points) that are used as inputs to knowledge-based prediction algorithms
* Collected 28374 data points distributed over entire globe at 217 AERONET sites during years 2005 and 2006
* Make use of RMSE between vector of target values and vector of corresponding predictions to assess AOD prediction accuracy
* Also report accuracy on domain specific measure fraction of successful predictions (FRAC) that penalizes errors on small AOD more than errors on large AOD
  + AOD prediction considered successful if absolute error is: 
* Define FRAC as: , where I = number of predictions that satisfy the above relation
* Primary benchmark for comparison is most recent version of MODIS deterministic algorithm called ‘C005’
  + Deterministic algorithms that retrieves AOD from MODIS observations rely on domain knowledge of aerosol properties and are based on lookup tables representing most common atmospheric conditions
* As a baseline statistical algorithm, used a NN trained to predict AERONET AOD from all MODIS attributes except location info and quality flag (hidden layer w/ 10 nodes and output layer w/ 1 node)
* For CRF model, consider first case when interaction potential does not exist (i.e. B=0)
  + NN and C005 predictions are inputs to CRF and each of 5 indicator functions belong to one of 5 world regions, and determined 10 ‘α’ parameters corresponding to C005 and NN predictions over these regions
* Over all regions, CRF achieved a better accuracy than either NN or C005 alone
  + Values of obtained ‘α’ params suggest to trust the NN more in North America while in Africa should trust the C005 more
  + Also, CRF improves domain-based accuracy measure FRAC
* Second case, check how much we should rely on NN and C005 over observations with different qualities
  + Partition data into 4 subsets and introduce 4 indicator functions to indicate belonging to each of these subsets and determine eight ‘α’ parameters corresponding to C005 and NN predictions over those subsets
* For all data qualities, CRF achieve better accuracy than either NN or C005 alone
  + Error of C005 decreases as data quality increases
  + Obtained ‘α’ values suggest to trust the NN more for low data quality set while for high data quality set should equally trust C005 and NN
  + FRAC also improved here by CRF
* Next consider case where interaction potential exists (i.e. B != 0)
  + NN and C005 predictions still inputs to CRF
  + Defined spatial and temporal neighbours as pairs of observations where temporal distance ‘temporalDist(i, j)’ is <60 days and spatial distance ‘spatialDist(i, j)’ is <100 km
* Taking into account the spatio-temporal correlation and comparing to the CRF with ‘B’=0 with world partitioned into 5 regions, we get better results globally here and over all regions separately except Africa where the two models are equally good
* Results suggest that the level of spatio-temporal correlation is different in different regions, and each region should have its own B
  + B estimated to be ~0.0459, which doesn’t indicate significant correlation, but still enough to improve single-output based predictions (where outputs are independent, i.e. B=0)
* Slightly higher B value learned, however, when data was partitioned based on quality rather than by regions, and improves final prediction
* Proposed method combines the outputs of a powerful non-linear regression tool such as a NN by incorporating a variety of correlated knowledge sources into single prediction model
  + Presented model can be applied to any regression application where there is a need for knowledge integration and exploration of structure in outputs

**Significant Points and Takeaways from Paper**

* Paper proposes CRF probabilistic model for structured regression that uses multiple non-structured predictors as features
  + Construct features as squared prediction errors and show that this results in a Gaussian predictor
  + Learning becomes a convex optimisation problem leading to a global solution for a set of parameters
  + Inference can be conveniently conducted through matrix computation
* Goal is to combine the two approaches (domain + statistical) so that in some regions where we know determ algorithm performs better, we rely more on deterministic model than on a statistical method, and vice versa
* Aerosol data is characterized by strong spatial and temporal dependencies that CRF are able to exploit by defining interactions among outputs using feature functions
* Use of features to define model is convenient because it allows us to include arbitrary properties of input-output pairs into compatibility measure; this way, any potentially relevant feature could be included to the model because parameter estimation automatically determines actual relevance by feature weighting
* Association potential function  associates single output ‘yi’ with all input data ‘x’ by real-valued function with ‘α’ being a K-dimensional set of params (doesn’t impose independency relations for ‘xi’)
* To model interactions among outputs, real-valued function called interaction potential  is used, where ‘B’ is L-dimensional set of params (represents relation between 2 outputs, e.g. in time or space)
* CRF models conditional distribution ‘P(y|x)’, y=(y1, …, yN) according to associated graphical structure
* 
* Most important diff w/ real-val targets is that the norm fun ‘Z’ is integral (Z must be integrable to be feasible model) + learning task is to choose values of parameters ‘α’ and ‘B’ to maximise conditional log-likelihood
* Condition for ‘P(y|x)’ to be multivariate Gaussian is that normalization function ‘Z’ is finite (i.e. if covariance matrix  is pos semi-definite), hence when learn params have to constrain covar matrix to be semi-definite
* Learning is thus a constrained optimisation problem because we need to guarantee that ‘ak’>0 and ‘B’>0
* Assoc and inter potentials can be defined as linear combins of a set of fixed features in terms of ‘α’ and ‘B’
  + 
* To model association potential (i.e. dependency between preds + target AOD), introduce 2 feature functions
  + , 
* To model interaction potential, introduces a feature function:
  + , ‘wij’ > 0 representing measure of spatio-tempor prox between data
* Learned params ‘α’ of linear combination of these features
  + 
* Resulting CRF model is thus: 
* Above equation can now be represented as a multivariate Gaussian distribution
* ‘C005’ determ algo used as benchmark; relies on domain knowledge of aerosol properties and lookup tables
* NN statist algo used as benchmark; rained to predict AERONET AOD from all MODIS attributes except a few
* NN and C005 predictions are inputs to CRF and each of 5 indicator functions belong to one of 5 world regions, and determined 10 ‘α’ parameters corresponding to C005 and NN predictions over these regions
* Over all regions, CRF achieved a better accuracy than either NN or C005 alone
* Proposed method combines the outputs of a powerful non-linear regression tool such as a NN by incorporating a variety of correlated knowledge sources into single prediction model
  + Presented model can be applied to any regression application where there is a need for knowledge integration and exploration of structure in outputs