7. EXPLORATION OF FOCAL AREAS

In addition to working with the 24 large GBRMPA regions and water bodies, it is possible to define very specific spatial and temporal domains that might represent areas of greater focus. For example, it might be of interest to model water quality patterns in a defined area proximal to a source of river discharge as part of an exploration into water quality responses to catchment outcomes.

Small spatial domains also presents an oppotunity to explore data assimilation options. The current project has access to four streams of water quality data (discrete AIMS niskin samples, AIMS FLNTU data, Satellite remote sensing and eReefs modelled data). Assimilating eReefs data (4km resolution) and Satellite data (1km resolution) as presented in the eReefs model data represents substantial computational overheads as a result of their high dimensionality. Whilst the discrete AIMS niskin sample is substantially more sparse, it does nonetheless present its own challenges when it comes to assimilation (see below).

We have three choices for combining the discrete AIMS niskin sample data with the eReefs assimilated model data:

1. aggregate together the average discrete (Niskin) sample and the average eReefs data or indices.
2. assimilate via an Ensemble Kalman Filter similar to the eReefs/Satellite data assimilation
3. define a Gaussian Process that incorporates both the discrete AIMS niskin data and eReefs assimilated data
4. assimilate via Fixed Rank Kriging

As a motivating example, we will use the discrete AIMS niskin and eReefs model data surrounding a single Dry Tropics Midshelf AIMS MMP site (Yongala). Yongala is a deep water site and thus the eReefs and discrete AIMS niskin samples are likely to have been collected across a relatively homogeneous bathymetry. Initial discussions will focus only on data from a single day (25/03/2017). The spatial configuration of eReefs observations relative to the AIMS MMP Yongala niskin sampling location is displayed in Figure 108.

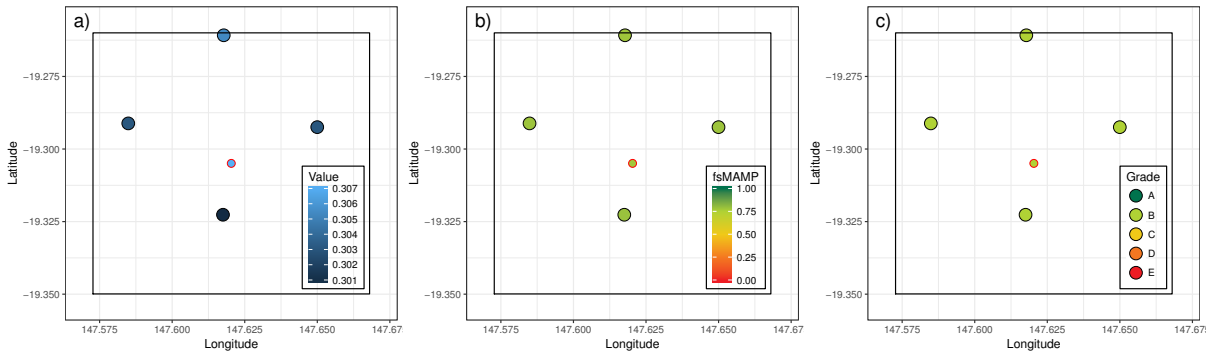

Figure 108: Spatial distribution of eReefs observation locations within 5km of the Yongala AIMS MMP niskin sampling location (point with red outline). Observations represent a) Chlorophyll-a values and associated b) fsMAMP indices and c) Grades (Uniform control chart) for 25/03/2015.

Importantly, although the AIMS niskin sample is located geographically roughly in the middle of the eReef locations, its Chlorophyll-a value (and fsMAMP index) is higher (and lower) than the surrounding eReefs values. Although this is only subtle in this example, it will be drawn upon when discussing aggregation options.

The fact that the observed AIMS niskin Chlorophyll-a sample collected on 25/03/2015 is higher than the surrounding eReefs estimates might suggest that either or both observation sets are only representative of limited scales. More specficially, it is likely that whilst the AIMS niskin samples only acurately reflect very local conditions, the 4km eReefs data are only likely to be reflective of broad larger scale conditions[21].

The above situation is likely to be exacerbated in highly heterogeneous seascapes. AIMS niskin samples are typically collected in close proximity to coral reefs where the general hydrology and input process might be

[21] The eReefs observations represent average modelled conditions within a 4x4km square cell, and therefore whilst potentially broadly reflective of large scale conditions, may not actually be an acccurate reflection of anywhere in that 4x4km cell

substantially different to the surrounding deeper water. By contrast, the eReefs model is known to be less reliable in shallow water. Thus, in areas that are heterogeneous with respect to bathymetry and hydrology, the AIMS niskin observations are likely to be representative of only the immediate vicinity (with very similar hydrology etc), whereas the eReefs observations might represent 'average' conditions that are only appropriate when considered on relatively large scales. The 4km resolution of eReefs model is unlikely to present adequate granuality in areas that are heterogeneous with respect to bathymetry and hydrology

Hence, the scale incompatibilities are likely to limit the ability to combine these two sources of data in a meaningful and reliable manner.

It is also possible that the accuracy of the two sources differ. Unfortunately, in the absence of a 'truth' this is difficult to assess. Nevertheless, since the eReefs data are indirect measures, it is possible that they are not as accurate as the AIMS niskin observations. If we had co-located observations (observations collected at the same locations and times from each source), we could attempt to align or calibrate the sources to one another. However, it is not possible to perform such alignments when data are not co-located and there is suspected differences in their spatial representation envelopes.

7.1 Simple aggregation

If we initially ignore all temporal aspects of the data and focus on the single day (25/03/2015), we could aggregate together the single discrete AIMS Niskin sample observation with the average of the four eReefs observations to yeild a single Chlorophyll-a estimate for the Yongala focal area (see Figure 109a). Alternatively, we could aggregate Chlorophyll-a indicies (see Figure 109b-c).

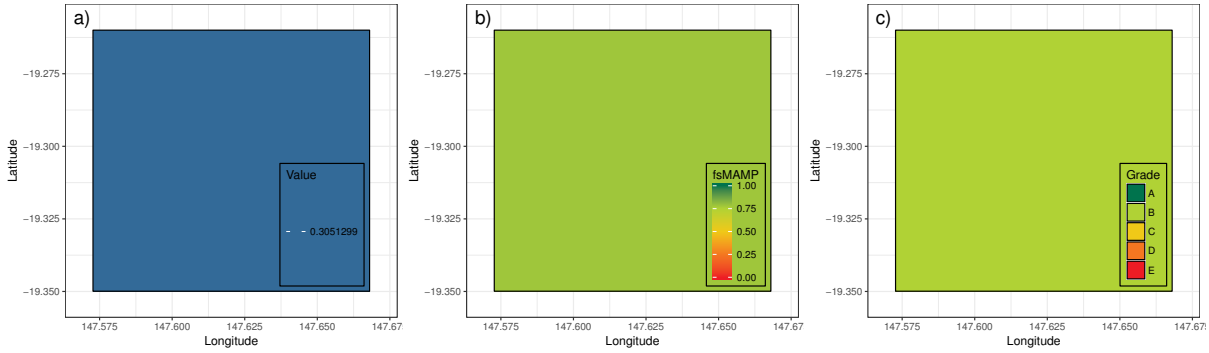

Figure 109: Yongala focal area aggregated a) Chlorophyll-a values and associated b) fsMAMP indices and c) Grades (Uniform control chart) for 25/03/2015.

Critically, this technique does assume that the single discrete AIMS niskin sample is representative of the entire spatial domain of the Yongala focal area. That is, we assume that the focal area mean is equal to this single point estimate. As previously discussed, this is likely to be an unrealistic expectation. We currently do not have any information on the spatial envelope represented by discrete samples. It is highly likely that the discrete samples are spatially biased (unrepresentative of the broader area as they are typically designed to sample reefs rather than the general water body. Rather it is likely that the discrete sample only represent the emmediate vacinity and uncertainty should decline with increasing distance. That said, the form to which certainty (representation) declines is completely unknow making it impossible to incorporate.

Furthermore, for the purpose of propagating uncertainty, the spatial uncertainty associated with the AIMS niskin sample is assumed to remain constant throughout this focal area. That is, out confidence in the focal mean is informed purely in our confidence in the single observation and that there is no additional loss of confidence associated with increasing distance from the sampling location. Obviously, it is highly unlikely that the reliability of the estimate will remain constant. The same is true for eReefs data, although it is likely to be less of an issue due to the greater sample size and spatial extent.

7.2 Ensemble Kalman Filter data assimilation

This is the approach used to assimilate the Satellite data into the eReefs model. Data Assimilation (DA) is a technique with forcasting and reanalysis, the later of which involves conditioning estimates of state on multiple

sources of data. For example, high density modelled data based on thermodynamics and gas laws might be 'calibrated' or orgmented by data observed at weather stations. The Kalman filter estimates state as the joint probability distribution ($p(x|y)$) which according to Bayes rule is proportional to the prior probaility ($p(x)$) multiplied by the probability (likelihood) of the observational data ($p(y|x)$). The simple Kalman filter provides algebraic expressions that describe the transition of state mean and covariance over time assuming all probability density functions are Gaussian and the transition is linear. If we say we have a prior belief that the state (x) has a mean of μ and covariance of Q and that the data (d) have an expected value of Hx and covariance of R, it can be shown that the posterior mean ($\hat{\mu}$) and covariance $\hat{Q}$ are:

$$\hat{\mu} = \mu + K(d - Hx), \hat{Q} = (I - KH)Q$$

where K (the Kalman gain) is:

$$K = QH^T(HQH^T + R)^{-1}$$

Unfortunately, as the domain of x increases (higher dimensionality), the covariance becomes prohibitatively large. If however, the state space (x) is broken up into a series of states (each perhaps representing a small subset (or ensemble) over time/space), we can replace Q with C (the sample covariance).

In either case, we must have estimates of both C and R. Whilst we can obtain estimates of C, estimates of R are not possible. If we only have a single discrete value within a higher-dimensional model domain, then we have no way of estimating R. Furthermore, even in larger focal areas that might contain multiple discrete samples, the samples are too spread out both spatially and temporally to be able to estimate R with any accuracy or reliability. For example, whilst the samples are typically separated in space by 10's of kilometers and months in time, water samples are likely to vary over the scale of meters and hours.

7.3 Gaussian Processes

A Gaussian distribution represents the distribution of observations that are themselves the result of an infinite number of influences (or processes). They are widely used to represent the distribution of residuals (unexplained component) when modelling data as it is often assumed that the unexplained component is due to a huge number of additional, unmeasured influences. In traditional linear modelling, we assume that not only are the residuals normally (Gaussian) distributed, we also assume that they are independent (not spatially or temporally correlated) and equally varied around 0.

$$\varepsilon_i \sim \mathcal{N}(0, \Sigma) \qquad \Sigma = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & \vdots \\ \vdots & \dots & \sigma^2 & \vdots \\ 0 & \dots & \dots & \sigma^2 \end{pmatrix}$$

Similarly, rather than express the stochastic elements as a vector of residuals drawn from a normal distribution, we can model the observed data as a multivariate normal (Gaussian) distribution. In this case, we are assuming that each of the observations is drawn from a multivariate normal distribution with different means and covariances).

This same argument could be extended to describe the distribution from which functions are drawn. Observed data are the result of the sum of an in infinite number of processes (including measurement error). Many of these processes vary over space and time such that sampling units that are closer together in space and time tend to be more similar to one another than they are to more distant units.

$$y_i \sim \mathcal{MVN}(M, C)$$

A Gaussian Process is largely defined by the covariance matrix ($k(x, x')$). Actually k is referred to as the **kernel**. We can define any covariance (kernel) function provided it is semi-definite - essentially that it is a symmetrical matrix.

A few of the popular kernels are described in the following Table 17.

Table 17: Simple Gaussian Process kernel functions

Kernel	Function
Linear	$k(x, x') = \sigma_f^2 xx'$
Squared exponential	$k(x, x') = \sigma_f^2 exp\left[\frac{-(x-x')^2}{l^2}\right]$
Periodic	$k(x, x') = \sigma_f^2 exp\left[\frac{-2\sin^2(\pi(x-x')/p}{l^2}\right]$
Periodic exponential	$k(x, x') = \sigma_f^2 exp\left[\frac{-2\sin^2(\pi(x-x')/p}{l_1^2}\right] exp\left[\frac{-(x-x')^2}{l_2^2}\right]$
Matern	$k(x, x') = \sigma_f^2 \frac{1}{\Gamma(v)2^{v-1}}\left[\frac{\sqrt{2v}\lvert(x-x')\rvert}{l}\right]^v K_v\left[\frac{\sqrt{2v}\lvert(x-x')\rvert}{l}\right]$

In Table 17, x and x' are vectors of the X variable. x' just indicates a transposed version of the vector. Hence $(x - x')$ indicates the difference (distance) between each pair of x values (they are squared so that they are all positive). When two points are similar, $k(x, x')$ approaches 1 (perfect correlation). Smoothing is based on neighbours exerting influence on one another (being correlated). When two points are very distant $k(x, x')$ approaches 0. The l are length scale parameters that determines the degree of contagion - that is, they determine the rate that the influence of points deteriorates with distance.

Assuming that the covariance pattern defined by the GP parameters (e.g. σ_f^2 and l) and observation space reliably reflects the underlying processes, the same parameters can be applied to yield a covariance structures for predicting mean and variance across a noval (yet overlapping) space. Specifically, if the covariance across the observed space is K_{oo}, the covariance between observed and prediction space is K_{op} and the coavariance across prediction space is K_{pp}, then the mean and variance for predicted values are:

$$\bar{y}_p = K_{op}(K_{oo} + \sigma_o^2 I)^{-1} K_{op}^T y_o$$

and

$$var(y_p) = K_{pp} - K_{op}(K_{oo} + \sigma_o^2 I)^{-1} K_{op}^T$$

where σ_o^2 is the estimated variance (uncertainty) in the observations, I is an identity matrix of equivalent dimensionality to K_{oo} and K_{op}^T is the transpose of K_{op}.

Gaussian Processes could be used to fit smooth multidimensional smoothers separate over each source so as to estimate parameters and uncertainty at any granulariy. Whilst this might be appropriate for the eReefs data, it is not possible to build a reasonable gaussian process via a single point without external estimates of the covariance over functions (σ_f^2) and the length (wiggliness) of the smoother.

Normally a Gaussian Process is applied to a single source for the purpose of kriging (smoothing). Nevertheless, it could be argued that there are a single set of underlying processes driving spatio-temporal patterns of water quality (e.g. l and σ_f^2) and that the multiple sources (AIMS niskin and eReefs) represent alternative ways to sample observations from those processes. Ideally, any differences between the sources should purely be differences in accuracy and uncertainty. If this is the case, rather than assume all observations are associated with the same σ_o^2, we could asssociate one variance to the AIMS niskin observations (σ_n^2) and another to the eReefs observations (σ_e^2).

Figure 110 illustrates a squared exponential Gaussian Processes with different parameter values applied to a single dimension (Latitude) of the 25/03/2015 Yongala focal area data. In each case, the variability (uncertainty) of the AIMS niskin observations was defined as 10 times lower than than of the eReefs observations. Values of σ_f^2 and l were chosen to represent specific sets of scenarios. For example, lower σ_f^2 imposes a lower maximum covariance

and a lower l dictates are more rapid decline in the autocorrelation over distance. Whilst it is possible to apply these functions in an optimizing framework so as to allow the data to determine the most appropriate values for σ_f^2 and l, σ_n^2 and σ_e^2 must be supplied based on external estimates.

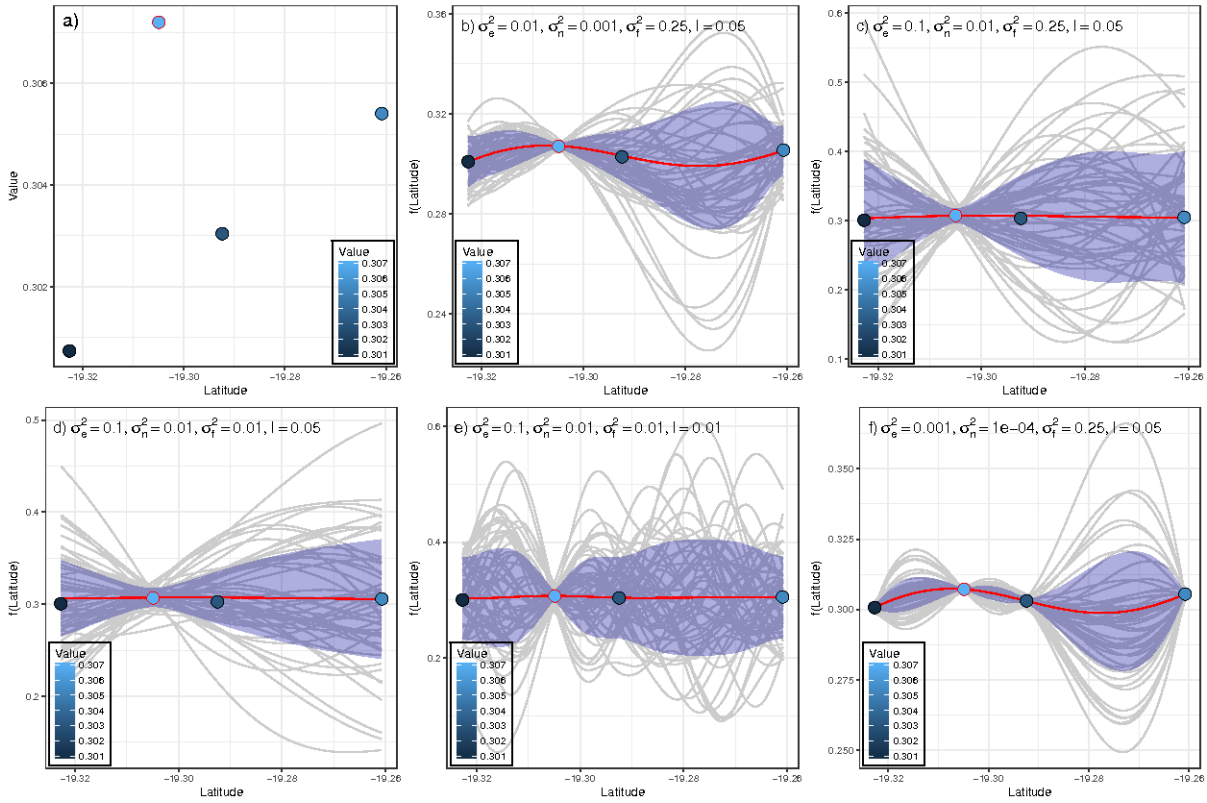

Figure 110: Illustration of data assimilation via squared exponential Gaussian process applied to a single dimension (Latitude) for the 25/03/2015 Yongala focal area a) Raw Chlorophyll-a values and b-e) different Gaussian Process parameters.

Similar to the Kalman Filter, high dimensionality incurs substantial covariance size increases. Every one additional observation results in a doubling of the covariance matrix and a tripling of memory to invert this the covariance matrix. Hence, practical applications employ either ensemble-like approaches or more commonly, sparse covariance matrices[22] to reduce the imposition of dimensionality.

Addition of a temporal dimension substantially increases the complexity of the problem. Not only does the covariance structures have to account for variability and autocorrelation length over space, it also has to reflect patterns of variability over time. Importantly, it is not just how isolated spatial points change over time. Temporal autocorrelation also occurs between neighbouring points.

7.4 Fixed Rank Kriging

Fixed Rank Griging (FRK) is a spatio-temporal modelling and prediction framework in which spatially/temporally correlated random processes are decomposed via linear combinations of basis functions (Φ) along with associated fine-scale variation (ν) (Cressie and Johannesson, 2008).

$$Y = X\beta + \Phi\alpha + \nu$$

The use of relatively small numbers of basis functions permits substantial dimensionality reductions that offers a scalable solution for very large data sets. Moreover, the framework facilitates differing spatial support hence allowing some capacity for the 'fusion' of multiple sources with different footprints.

[22] Sparse matrices acknowledge that covariance will decline over time and distance and at some distance, the covariance will effectively be zero.

Varying footprints are accomodated by arranging the point-referenced data into grids, the granularity of which is proportional to the footprint or extent of support. For example, the AIMS niskin data and eReefs modelled data could be descretized into a small and set of larger grid squares (see Figure 111b - pale red and blue squares respectively). Whilst the footprint size for the eReefs modelled data was based on the cell grid onto which the model is projected, the AIMS niskin footprint was set to an arbitrarily (smaller) value to illustrate varying degrees of support.

The full spatio-temporal domain is also discretized into a regular grid of smaller cells called *basic areal units* (BAU) which represent the smallest modelling and prediction unit. In this example, we have discretized the spatial domain by hexagonal cells 0.01 degrees longitude by 0.01 degrees latitude (see Figure 111b - black hexagons). Within the model, varying support is then based on the intersection of the square footprints with the BAUs.

For this example, we have elected to define two regularly spaced basis functions based on Matern covariance (smoothing parameter of 1.5) to be used in the decomposition of spatio-temporal processes (see Figure 111c). The multiple resolutions provide a mechanism for esimating the scale of spatio-temporal autocorrelation (however, ideally this requires a substantially larger grid of data than our example).

The basis function covariance matrices and fine-scale variance parameters are estimated via a expectation maximization (EM) algorithm and thereafter used to project predictions onto the scale of the BAU's (see Figure 111d). These predicted values have also been indexed via fsMAMP (see Figure 111e) and converted into Grades (see Figure 111f).

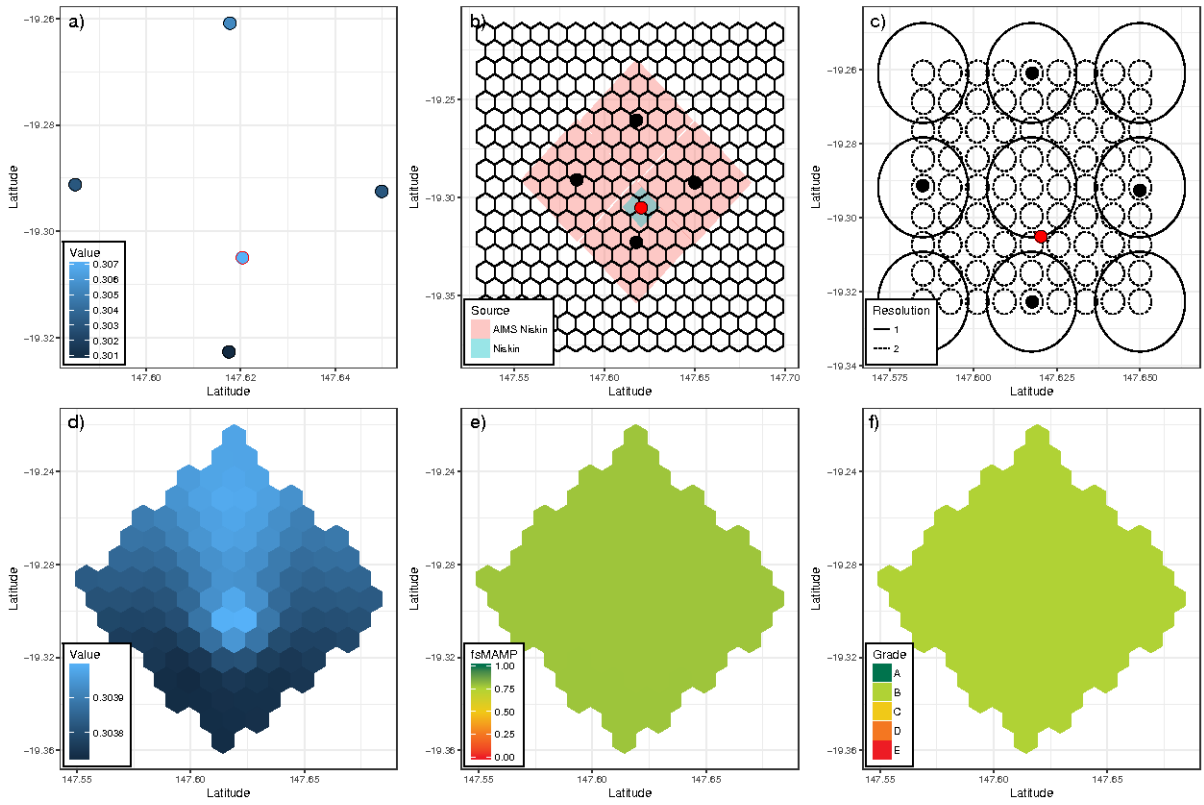

Figure 111: Illustration of data assimilation via Fixed Rank Kriging applied to spatial data for the 25/03/2015 Yongala focal area a) Raw Chlorophyll-a values (AIMS niskin: red symbol border, eReefs: black symbol border), b) dicretization of the spatial domain into a regular hexagonal grid and varying footprints (support) for AIMS niskin (blue) and eReefs (red), d) Matern basis functions of two resolutions, d) predicted values and associated e) fsMAMP indices and f) Grades (Uniform control chart) for 25/03/2015.

Figure 111 illustrates that whilst fixed rank kriging does offer an option for the assimilation (or fusion) of multiple data sets, in the absence of measurement error, it does assume that all observations are equally accurate. Figure 111d shows a bright spot associated with the higher AIMS niskin Chlorophyll-a value. It is important to reiterate that the extent of this bright spot is due to both the higher Chlorophyll-a observation of the AIMS niskin sample and the arbitrary size of the footprint. To be a meaningful fusion, reasonable estimates of the spatio-temporal

extent of representation of the AIMS niskin data will need to be obtained along with estimates of measurement error in both the AIMS niskin and eReefs modelled data.

Spatio-temporal basis functions can be constructed as the tensor product of spatial basis functions and similarly defined temporal basis functions. Measurement error (if known) can also be incorporated.

More recently, Nguyen et al. (2014) has proposed a data assimilation technique for big data that is essentially a blend of fixed rank kriging and Kalman filtering and looks to have some promise.