Data-driven behavioural characterization of dry-season groundwater-level variation in Maharashtra, India

Rahul Gokhale* and Milind Sohoni

Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Mumbai 400 076, India.
**Corresponding author. e-mail: rahulbg@cse.iitb.ac.in*

This paper looks at the crucial issue of dry-season groundwater-availability in the state of Maharashtra, India. We look at the two key hydro-climatological measurements which are used to implement groundwater policy in the state, viz., water levels in 5000+ observation wells across the state and aggregate rainfall data. We see that there is substantial variation in groundwater levels within and across the years in most wells. We argue that for a large number of these observation well locations, aggregate rainfall data is inadequate to model or to predict groundwater levels. For this, we use a novel random rainfall coefficient model for the purpose of modelling the effect of rainfall in a composite setting where extraction and changing land-use data is unknown. The observed high variance of this coefficient points to significant variations in groundwater levels, which may only be explained by unmeasured anthropogenic factors. Next, we see that the uncertainty in actual groundwater levels along with scarcity are two distinct features of groundwater availability and will elicit different behaviours from the typical user. Finally, we recommend that quantitative groundwater assessment protocols of the state should move to incorporating data from which extraction and land-use may be modelled. We believe this is one of the first studies where large spatio-temporal scale data gathered by state agencies have been analysed for scientific adequacy.

1. Introduction

Groundwater is the major source of freshwater for drinking and agriculture in most parts of India. In the state of Maharashtra, almost all groundwater is obtained from shallow to moderately deep unconfined hard-rock aquifers (see GSDA 2010). Due to this low groundwater-potential, groundwater availability in the state relies heavily on the yearly recharge from annual rainfall, more than 90% of which occurs during the monsoon season, from June to September. This recharge is the major source of freshwater during the dry-season from October to May. At the same time, in recent decades, increasing population and economic activity has led to a steep rise in the dry-season groundwater demand. This has made the correct estimation and predictability of the resource an important objective for the state administration.

In Maharashtra, groundwater is administered, at least for its technical aspects, by the Groundwater Surveys and Development Agency (GSDA), a technical agency of the Government of Maharashtra (GoM). The agency, through the Senior Geologist of every district, puts out a report in the month of October of every year on the groundwater outlook during the ensuing dry-season for that district. This forms an important input to the District Collector. A sample report for the district of Kolhapur and the dry-season during year 2013–2014 is available online, see GSDA (2013). The report is based on the water levels in a set of observation wells in

Keywords. Groundwater level variation; multilevel statistical regression; uncertain groundwater availability.

that district and the aggregate rainfalls reported at each *taluka* (administrative sub-block of district) place (see table on page 4 of the report). Our work looks at the phenomenon of variation of groundwater levels and the possibility of predicting them from such type of data and shows that a certain class of models perform poorly for that purpose.

This paper uses groundwater-level data recorded over the recent 2 to 4 decades from 5000+ observation wells located throughout Maharashtra and monitored by GSDA. On the average, there is roughly one well per 60 km^2, and furthermore, a dug-well is sampled four times every year and a bore-well, once every month. Thus, the data though wide in expanse, is spatio-temporally sparse for fractured hard-rock groundwater-regimes.

Considering such nature of the data, we do a temporal analysis and modelling of dry-season groundwater levels separately for each well, and the parameters so obtained are mapped for broad regional interpretation. These parameters express the impact of (i) rainfall and (ii) previous year's closing groundwater level, on the behaviour of dry-season groundwater levels. Besides mean groundwater levels, we regard variance in groundwater levels as an important attribute of the regional groundwater-regime, since both scarcity (i.e., deeper mean groundwater level) and uncertainty (i.e., high groundwater-level variance) separately impact the typical groundwater-user.

To estimate the variation of aggregate dry-season groundwater-availability across the years, we use an adaptation of variance components modelling, that estimates the *interannual* variance component of groundwater-level variation, separately from the *intra-(dry-)seasonal* variance. The intraseasonal variation is usually caused by location-specific events such as nearby extraction, canal operations, etc., which are, in principle, observable by the groundwater-user. It is the interannual variation which is the focus of this paper, for it impacts key economic decisions of the groundwater users as a community, such as crop mix or investments in water conservation technology over the years. Next, we formulate our *rain-based* model to estimate the dependence of aggregate dry-season groundwater levels on the amount of monsoon rainfall as well as the previous year's 'closing' groundwater level. The dependance on rainfall is modelled by treating the coefficient of rainfall as a random variable itself. This coefficient, as a multiplier, captures not only the composite effect of those hydrogeological parameters which do not change, but also anthropogenic factors such as cropping pattern, that modulate the net effect of rainfall across the years.

The results of our analysis show that there is substantial variation in actual groundwater levels from the best-fit models. Moreover, interannual variation is a significant component of the total variation. This is severe in the region under the Drought Prone Area Programme and in the Tapi–Purna river basin's alluvial deposition belt along the northern fringe of the state. Total monsoon rainfall and previous year's closing groundwater levels, together, are effective in predicting barely half of the interannual variation. This is demonstrated by the high variance in the rainfall effect. These findings imply that (i) for many users, groundwater is not only scarce, it is also uncertain, and (ii) groundwater assessment and outlook models will probably need to incorporate local socio-economic data in addition to the hydrogeological and climatological data that they currently do. These findings have direct implications for policy design and administration of groundwater in Maharashtra.

The paper is organized as follows. We point out the need for a data-driven analysis of long-term state-wide groundwater-level variation in section 1. Section 2 presents the datasets on groundwater levels and rainfall that are used for the work. Section 3 formulates the models, while their results are presented in section 4. Section 5 notes important points about our models and the implications of their results for groundwater-management and research. Conclusions are drawn in section 6.

2. Background and motivation

Research in the groundwater-sector may broadly be classified as (physical-)scientific, socio-economic or administrative (related to policy and its implementation). Much of the scientific literature about Maharashtra is in the form of location/area-specific case studies. Use of electrical resistivity tests to detect shallow groundwater sources has been attempted by Bose and Ramakrishna (1978). Kulkarni *et al.* (2004) propose local groundwater-demand management based on aquifer characteristics like diffusivity. Rokade *et al.* (2007) have used remote sensing data and GIS to map groundwater-potential of a watershed. Modeling and evaluation of groundwater-harvesting structures has been attempted by Gore *et al.* (1998). There are also some studies like those by Pawar *et al.* (1998) and by Kaplay and Patode (2004), that look at the quality aspect of groundwater and evaluate the pollution caused by industrial effluents. Groundwater-resource assessment in the Koyna river basin has been attempted by Naik and Awasthi (2003). Deccan Traps constitute more than 80% of Maharashtra's geology. Groundwater-related geological assessment of Deccan Traps was conducted by Adyalkar and Mani (1971). Quantitative aquifer characteristics of Deccan Traps have been studied by Deolankar (1980), while Kulkarni *et al.* (2000) propose a hydrogeological framework for the Deccan Traps.

In the socio-economic and administrative realm, Sekhri (2014) points out the importance of groundwater and groundwater levels in the socio-dynamics of rural areas. The impact of policy reforms on the sustainability of groundwater-resource has also been studied by Sekhri (2012). Banerji *et al.* (2012) demonstrate a case of a north Indian village, where village-level social contracts result into spatially efficient groundwater-allocation. Based on simulations of such socio-economic set-up, they infer that power sector reforms *vis-a-vis* groundwater, can lead to higher-level economic equilibrium with better farm-yields. Specific to Maharashtra, other studies like the analysis of the implementation process of groundwater-legislation in Maharashtra by Phansalkar and Kher (2006) focus on administrative aspects. A review of the groundwater-sector in the Deccan Traps geology of Maharashtra, by Limaye (2010), addresses the developments in the three spheres of hydrogeology, economics and administration of groundwater-use.

Data-driven modelling has been increasingly applied in situations that are phenomenolgically complex, such as weather predictions, and in situations that are complex interactions between natural and human factors. Groundwater is increasingly one such area where data-driven models have been profitably used. For example, artificial neural networks are used for prediction by Lallahema *et al.* (2005). Kholghi and Hosseini (2009) compare the utility of adaptive network-based fuzzy inference systems and ordinary kriging in predicting groundwater levels. Giustolisi and Savic (2009) have considered a combination of polynomial regression and genetic algorithms for groundwater-level predictions. Shirmohammadi *et al.* (2013) have studied the use of several data-driven techniques for predicting groundwater levels. A framework for enhancing physically-based predictions of groundwater levels using various data-driven models has been developed by Demissie (2008). Few other studies like the one by Passarella *et al.* (2013) have used optimization techniques, deterministic as well as stochastic, for an optimal redesign of groundwater-monitoring networks.

We hope that our study contributes in two directions. On the data-driven modelling front, ours is perhaps the first multilevel statistical regression model for seasonal groundwater variations, and that which works on the region-wide, yet sparse, datasets presented by government agencies. For Maharashtra, despite the serious issues of seasonal and regional scarcity, no such modelling has been considered. Secondly, our random coefficient formulation for rainfall parameter also provides an interesting interpretation about *net* rainfall-effect on groundwater-availability. This coefficient may serve as a composite metric of the health of a groundwater system.

3. Datasets

The two spatio-temporal datasets that we use in this paper are the 2–4 decade long state-wide groundwater level dataset for Maharashtra and a similar daily rainfall dataset. These datasets are described below.

3.1 *Groundwater-level dataset from GSDA*

GSDA began monitoring groundwater levels in Maharashtra starting with mere four observation wells (also called ‘sites’) before 1970. The number of sites gradually rose, and is 5383 as available in our dataset (year 2011). Among these, 4260 are dugwells, 1108 borewells, 14 tubewells and 1 dug-cum-borewell (see figure 1).

Observations in dugwells are usually recorded once in January, March, May and October each, while those in other types of wells are usually recorded once every month. Observations are recorded as the date of observation and the groundwater level in metres below ground-level (mbgl). The period of four months from June to September every year is considered as the monsoon season. The first observation of the dry-season is often recorded in the last 5 to 10 days of September. So we mark the period from September 21st (instead of October 1st) to May 31st as the dry-season and the groundwater levels recorded in this period are considered as the dry-season groundwater levels. The May groundwater level (the latest one, when more than one observation has been recorded in May for some reason), has been considered as the pre-monsoon groundwater level. For convenience in modelling, we delineate years as hydrological years, June to May.

Each well also carries a watershed label, elevation (above mean sea level), and the location, i.e., latitude and longitude of the well. It should be informed that, classification of sites by their watersheds or elevations did not yield any significant insights and will not be used in the present analysis. Moreover, each watershed has an attribute called ‘average depth to bottom of aquifer’, which is perhaps an aggregate of various wells within the watershed. This too was not found useful or significant. The location

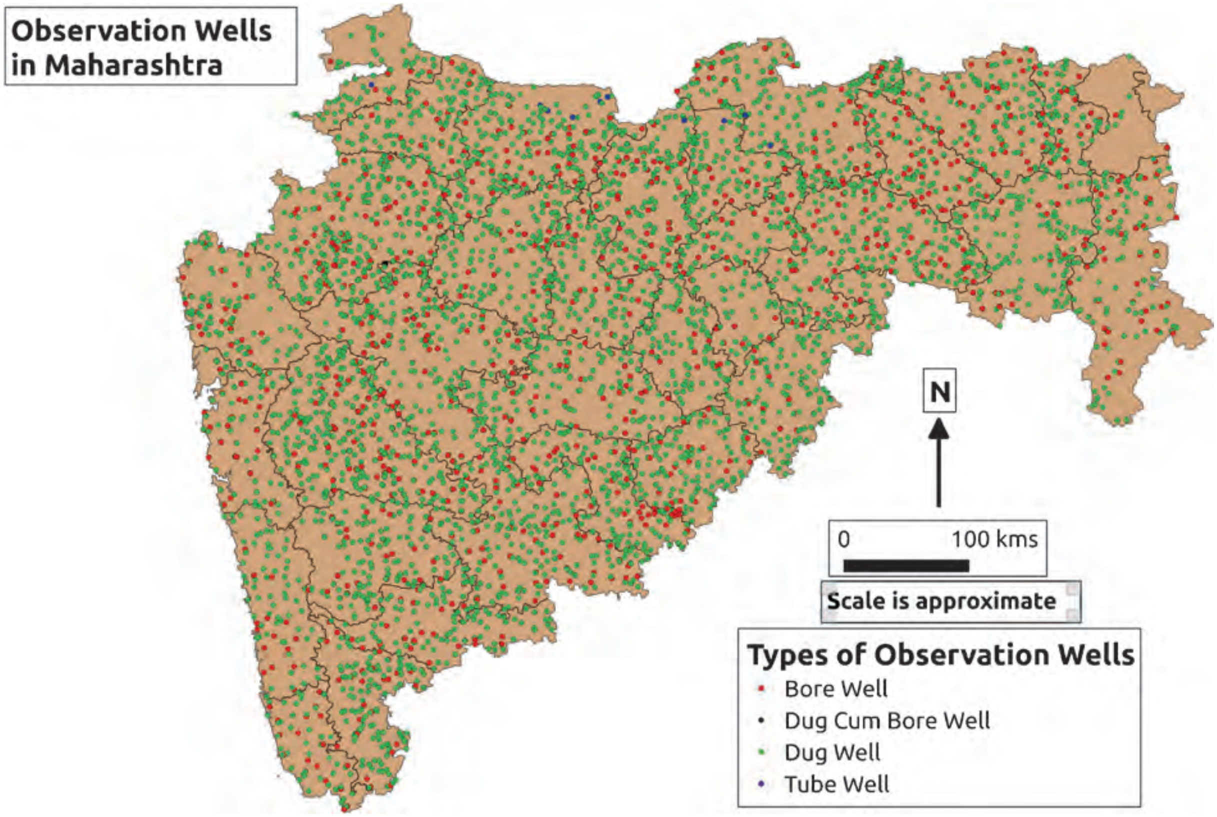

Figure 1. Location of observation wells in Maharashtra.

information is used for mapping and broad spatial interpretation.

3.1.1 *Dry-well observations and other data irregularities*

When an observation well is dry, its reading is recorded (by GSDA) to be same as the well-depth, as a convention. The aquifer depth at well locations was not available during analysis. Thus, the validity of treating a dry-well reading as a groundwater level and using it as it is, for modelling, needs care. After considering various options for incorporating dry-well readings, and experimentation (like Maximum Likelihood estimation under data censoring), we have chosen to admit these readings for those wells where the fraction of such readings (in the total dry-season readings of the well) is not large. As far as the estimation of groundwater-level variation is concerned, this will underestimate the total variance. Inferences based on 'high variance' will then only be conservative. There are 3157 wells in the dataset, which have at least one dry-well reading in the dry-season of some year. However, only 292 of these are such that more than 25% of their total dry-season readings are dry-well readings. Even if these observation wells are discarded from analysis, there is no visible lack of observation wells in any region of Maharashtra. So, these 292 wells have been discarded from our analyis.

There is considerable irregularity about the exact date of observation as well as the monitoring schedule. For example, observations to be taken at the beginning of October are sometimes recorded in November. Also, in the early years of monitoring (1970s), only post-monsoon and pre-monsoon groundwater levels used to be recorded. There are also gaps (of many years) in the monitoring of many sites. Such features of the dataset make it less amenable to standard time-series analyses, but do not hinder the use of regression methods based on curve-fitting, as are used here.

3.2 *Rainfall dataset*

The National Climate Centre of India Meteorological Department (IMD) at Pune, India, has developed a rainfall dataset containing daily rainfall values for points of a $0.5^\circ \times 0.5^\circ$(lat., long.) grid over India (see Rajeevan and Bhate 2009). Of this, the rainfall data for 30 years from 1975 to 2004 for the grid-points which overlay Maharashtra state has been used for this work.

There are roughly 30 to 60 wells in a single cell of the $0.5^\circ \times 0.5^\circ$ (lat., long.) grid. Rainfall at observation well locations is not measured. Rainfall values at taluka stations, as also used by GSDA, have gaps that need processing and may not be the nearest station to the observation well. The IMD rainfall dataset already incorporates such processing and amalgamation and was found more suitable for this study. To obtain rainfall values corresponding to the observation wells, we have interpolated using rainfall values at the grid's vertices. The intepolation method used is modified Shepherd's method (see Shephard 1968), also called Inverse Distance Weighting (IDW) method, with range of influence bounded to 0.5 degrees (distance on the (lat., long.) grid) and with power parameter 1. Since the distances from the grid vertices of any two observation wells differ, each well within a grid-cell gets assigned a distinct rainfall series.

4. Modelling

As already noted, models are developed separately for each well and their results are mapped to conduct a broad regional comparison. In our models, for any given well, l denotes the groundwater-level variable, y denotes the variable for (hydrological-) the year and t denotes the variable for the day of the year. A groundwater-level observation is thus represented by (y, t, l). We develop a sequence of three regression models to progressively refine the understanding of groundwater-level variation. The statistical formulation of these models is presented below.

4.1 *Expected dry-season trend model*

Our baseline model for each well is the *Expected trend* model. It explains the expected groundwater level l on any given day, t, of dry-season as:

$$l_1(y,t) = a + bt + \gamma_t \tag{1}$$

where γ_t is random error on day t. Following standard practice in statistical regression, we assume the errors γ_t to be Identically and Independently Distributed (IID) as $\mathcal{N}(0, \sigma_\gamma^2)$. Note that l_1 ignores y to estimate the expected trend during any dry-season. This model forms the base-line for the total variance (beyond the expected dry-season discharge bt) seen by a general user in the groundwater-regime.

4.2 *Variance components model*

Our second model refines the description as shown below.

$$l_2(y,t) = a + bt + \alpha_y + \delta_{y,t} \tag{2}$$

where the random deviations α_y represent interannual variation, and the quantities $\delta_{y,t}$ represent intraseasonal variation, see figure 2.

The quantities α_y are assumed IID as $\mathcal{N}(0, \sigma_\alpha^2)$, and $\delta_{y,t}$ are assumed IID as $\mathcal{N}(0, \sigma_\delta^2)$. Random variations at the two time-scales are considered independent of each other, so that $\mathrm{var}(l_2(y,t)) = \sigma_\alpha^2 + \sigma_\delta^2$. The total variance of groundwater levels is thus split into interannual variance (σ_α^2) and intraseasonal variance (σ_δ^2). Technically, bt explains a part of the intraseasonal variation. However, being the

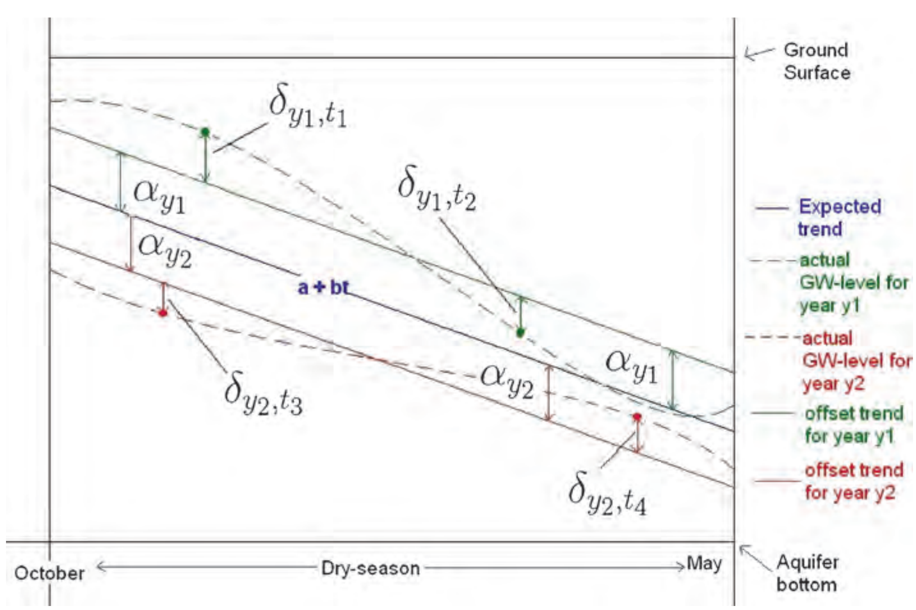

Figure 2. Conceptual model for dry-season groundwater-level variation.

dry-season, only the variation unexplained by this 'expected' discharge trend, is of interest. So, the regressor t is incorporated in this model and $\delta_{y,t}$ are made to account only for the residual intra-seasonal variation.

4.3 *Rain-based model*

Ideally, the interannual variation may be explained using two natural and widely acknowledged factors: the total monsoon rainfall, denoted by r and the pre-monsoon groundwater level, denoted by v. The former acts as a numeraire for monsoon rainfall, which is the primary input to the groundwater-regime, while v captures the status of groundwater-regime before this input occurs. The groundwater level may thus be predicted as $l(y,t) = a + bt + cr_y + dv_y$.

However, for the same rainfall, anthropogenic factors may lead to different groundwater-regime conditions across the years, thus modulating the *net* effect of rainfall in raising groundwater levels. This variation in rainfall-effect is modelled by treating the coefficient of r as a random variable itself. We do this by adding to the fixed-effect c, a random variable ρ. This random-coefficient model, termed as *rain-based* model, thus explains groundwater level as:

$$l_3(y,t) = a + bt + (c + \rho_y)r_y + dv_y + \epsilon_{y,t} \qquad (3)$$

where the prediction $a + bt + cr_y + dv_y$ mentioned above, forms the 'fixed part' of the model, while $\epsilon_{y,t} + \rho_y r_y$ forms the 'random (error) part'.

That variation in the aggregate dry-season groundwater-availability, which is not explained by $cr_y + dv_y$ is accounted by $\rho_y r_y$, while $\epsilon_{y,t}$ captures the residual intraseasonal variation on day t of year y. The random deviations ρ_y are assumed IID as $\mathcal{N}(0, \sigma_\rho^2)$, while the intraseasonal deviations $\epsilon_{y,t}$ are assumed IID as $\mathcal{N}(0, \sigma_\epsilon^2)$.

4.4 *Interpreting model parameters*

Our models attempt to describe the (temporal) behaviour of groundwater level variation. Their parameters, thus provide a behavioural characterization of the surrounding groundwater-regime. The parameter b estimates the dry-season's typical *discharge trend*, i.e., the average daily discharge rate in the surrounding groundwater-regime during dry-season. Note that b represents the effective behaviour of groundwater levels, as different from the purely hydrogeological natural discharge behaviour. The variance parameters σ_α and σ_δ are clearly behavioural indices.

The parameters c and σ_ρ relate to the rainfall effect. The net effect of monsoon rainfall on dry-season groundwater-availability depends on various natural and anthropogenic hydrogeological processes like rainfall pattern, infiltration, evapotranspiration, natural discharge, artificial extraction, artificial recharge, etc., most of which are in turn influenced by other socio-economic activites such as cropping, irrigation, and land use. These activities have their own cycles and trends, and hence the net rainfall-effect varies across the years, necessitating the use of random variable ρ. The coefficient $(c + \rho_y)$ of r, thus, abstractly encapsulates the impact of all such processes which may alter the effect of r. The fixed part c of this coefficient, represents the average rainfall-effect and is thus an index of the average favourability of the groundwater-regime conditions. The random part ρ_y represents the deviation from this average during the year y. A higher value of its variance (σ_ρ^2), reduces the predictability of aggregate dry-season groundwater-availability.

The pre-monsoon groundwater level v is incorporated in the rain-based model, mainly to eliminate the bias resulting from the differences in the pre-monsoon status of groundwater-resource. The parameter d is thus, an estimate of the empirical dependence of aggregate dry-season groundwater levels on v.

4.5 *Estimation method used*

The expected trend model is estimated using the ordinary least-squares method which gives Maximum Likelihood Estimates (MLE) of the model parameters. Both the variance components model and the rain-based model are forms of multilevel regression models (also known as hierarchical regression models). Goldstein (1995) provides a detailed description and a framework to develop and estimate such models, chapter 2 of which explains all that we have used in this paper.

The Restricted Iterative Generalized Least-Squares (RIGLS) estimation algorithm, a variant of IGLS algorithm, has been used to obtain parameter estimates for our models. We use 'less than 0.0001 change in every parameter within 100 iterations' to infer convergence to estimates. RIGLS provides unbiased estimates of all model parameters and those for model's coefficients are also MLE. IGLS, especially the unbiased estimation in RIGLS, can produce (usually only marginally) negative estimate of a variance component whose actual value is neglegible or zero. The number of such cases are mentioned in our results. Using a standard method, also described in Goldstein (1995), we obtain the posterior estimates of the residuals in the multilevel models; viz., $\hat{\alpha}_y$, $\hat{\delta}_{y,t}$, $r_y\hat{\rho}_y$ and $\hat{\epsilon}_{y,t}$.

5. Results

The results of model estimation for the three models of the previous section are presented below in the same order. Maps of spatial distribution of parameter estimates are collected together for ease of comparison and common spatial patterns seen in most of these maps will be discussed after all model results are presented.

5.1 *Results for expected trend model*

Figure 3 shows the graphical plot of the model for an example well. Instead of restricting ourselves to linear trends, we have also considered trends that are higher degree polynomials in t, to check their utility for state-wide assessment. For example, for a degree 2 polynomial model, equation (1) would become $l_1(t) = a + bt + ct^2 + \gamma_t$.

Table 1 shows the average values of R^2 (coefficient of determination) and $\hat{\sigma}_\gamma$ (standard deviation) for these models, the average being taken over all the 5000+ modelled sites. In general, there are meagre improvements in the model quality with increasing degree, indicating that higher degree components have negligible contribution in the typical pattern of dry-season groundwater-level trend. Furthermore, the risk of overfitting increases with the degree of polynomial, especially for irregularly sampled data. Thus, we will consider only linear trend models for general state-wide assessment.

A rasterized map of the spatial distribution of values of $\hat{b}$ (estimate of b) is shown in figure 4. 5309 wells have $\hat{b} > 0$ (discharge), which is significant (at 10%) for 5002 of them. Interestingly, there are 52 wells that have $\hat{b} < 0$, but only 3 have it statistically significant.

5.2 *Variance components model results*

Based on our convergence criterion, RIGLS did not converge for 18 sites. All results discussed henceforth, will be for the sites for which RIGLS converged. Figure 5 shows the graphical plot of the variance components model for an example site. Each year's posterior estimate of level-2 residual (α_y) has been incorporated while plotting, thus displaying a separate trend-line for each year.

Table 1. *Quality of the expected dry-season trend model.*

Degree of trend	Avg. R^2	Avg. $\hat{\sigma}_\gamma$
1	0.42	2.21
2	0.44	2.20
3	0.45	2.20

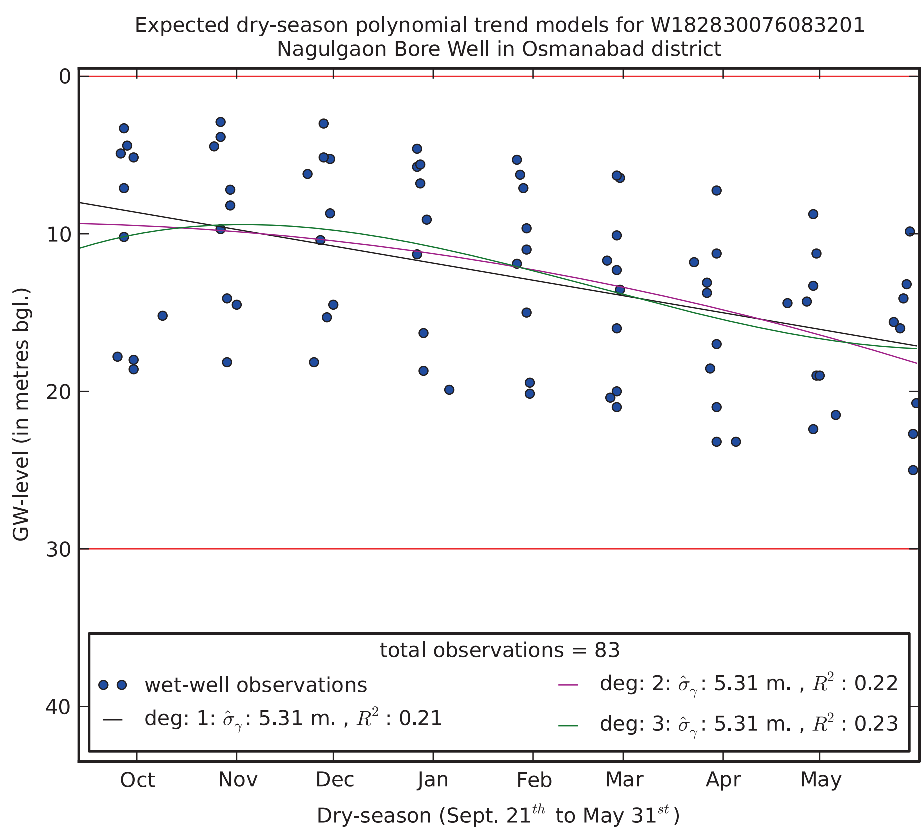

Figure 3. Example plot of expected polynomial trend models for dry-season groundwater levels.

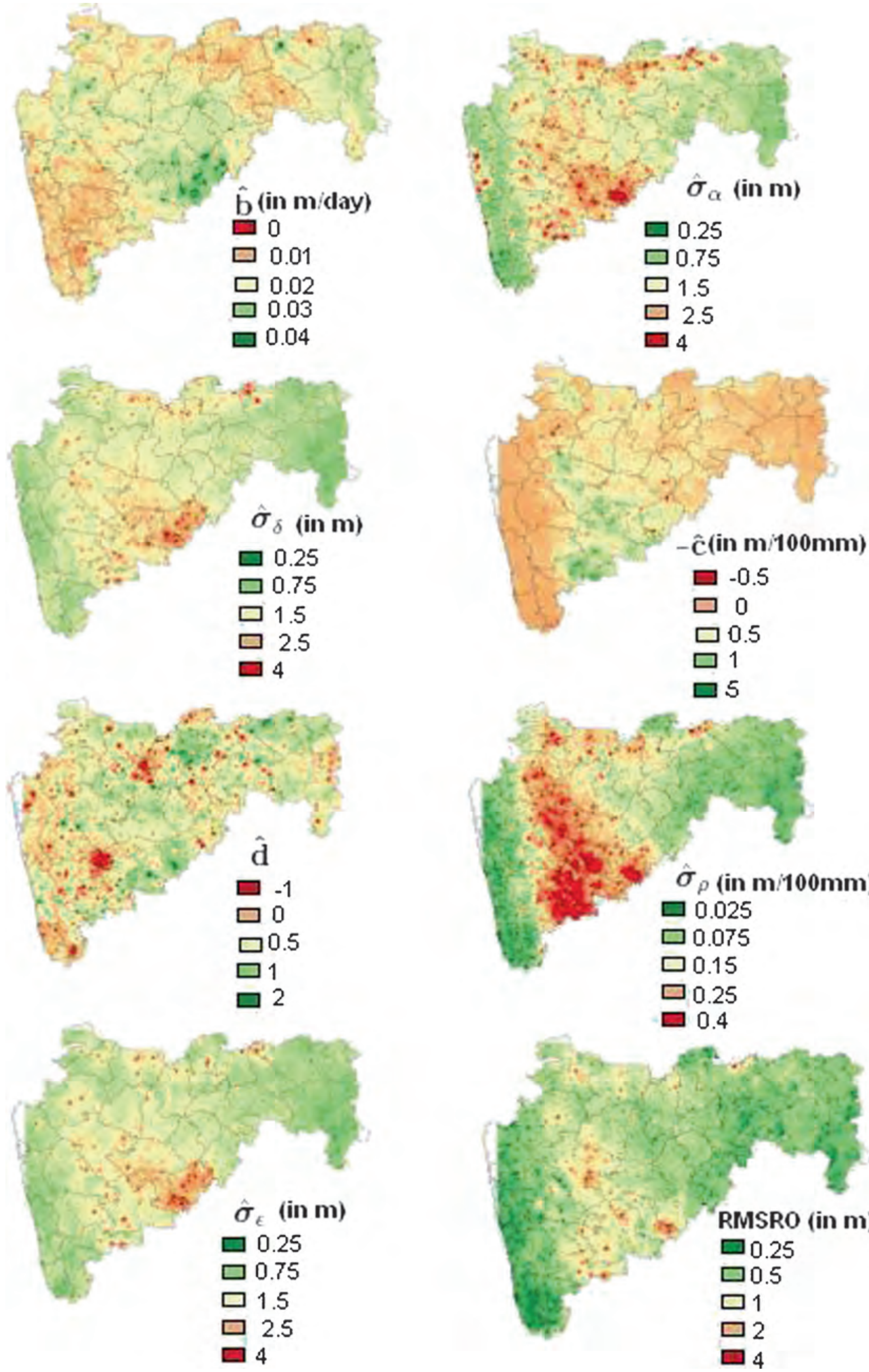

Figure 4. Spatial distribution of model parameters. Maps from top-left to bottom-right, row-wise: $\hat{b}$ (of expected trend model), $\hat{\sigma}_\alpha$, $\hat{\sigma}_\delta$, $-\hat{c}$, $\hat{d}$, $\hat{\sigma}_\rho$, $\hat{\sigma}_\epsilon$ and RMSRO.

Figure 4 includes maps of the spatial distributions of $\hat{\sigma}_\alpha$, the interannual variance estimate, and $\hat{\sigma}_\delta$, the intraseasonal variance estimate. The two maps show similar patterns, with the spatial distribution of $\hat{\sigma}_\alpha$ exhibiting sharper spatial patterns compared to those of $\hat{\sigma}_\delta$. 179 sites had a negative estimate of σ_α^2, while 59 sites had $\hat{\sigma}_\alpha > 10$ m. No site had a negative estimate of σ_δ^2, while 14 sites had $\hat{\sigma}_\delta > 10$ m.

5.2.1 *Severity of groundwater-level variances*

The validity of normality assumption about $\hat{\alpha}_y$s is tested using the Kolmogorov–Smirnov test for goodness-of-fit. Less than 10% of the sites show statistically significant (at 10%) lack-of-fit to $\mathcal{N}(0, \hat{\sigma}_\alpha^2)$. Now, with α_ys distributed as $\mathcal{N}(0, \sigma_\alpha^2)$, the (central-) interval such that α_y would lie in it with 90% probability, would roughly span $3.3\hat{\sigma}_\alpha$ metres. Among the 90+% sites, where $\mathcal{N}(0, \sigma_\alpha^2)$ distribution may be considered, 2360 wells have $\hat{\sigma}_\alpha > 1$ m, while 744 wells have it > 2 m. So, the 90% confidence interval regarding aggregate dry-season groundwater levels in these wells would be more than 3.3 and 6.6 m, respectively. In Maharashtra, the average aquifer thickness ranges between 10 and 25 m. We thus see that the probable (90% confidence-interval) interannual variation observed in large parts of Maharashtra is a substantial fraction of the maximum possible aggregate dry-season groundwater-availability. The intraseasonal variance further adds to the variation in groundwater-availability for any given day of the dry-season.

5.3 *Rain-based model results*

RIGLS did not converge for 91 sites. All results discussed henceforth, will be for the sites for which RIGLS converged. Figure 6 shows the graphical plot for an example site. It is a set of predicted trend lines, one per year. In figure 7, graphical plot for another site is shown, where level-2 residuals have also been included to plot dotted lines. Thus we have two lines for each year(y) in figure 7. The bold lines represent prediction model $\hat{l}(y,t) = \hat{a} + \hat{b}t + \hat{c}r_y + \hat{d}v_y$, as in figure 6, while the (corresponding colour) dotted lines represent the trend-line $\hat{l}(y,t) = (\hat{a} + \hat{b}t + \hat{c}r_y + \hat{d}v_y) + \hat{\rho}_y r_y$, arising from the (random) deviation $\hat{\rho}_y$ in the effect of r in the year y.

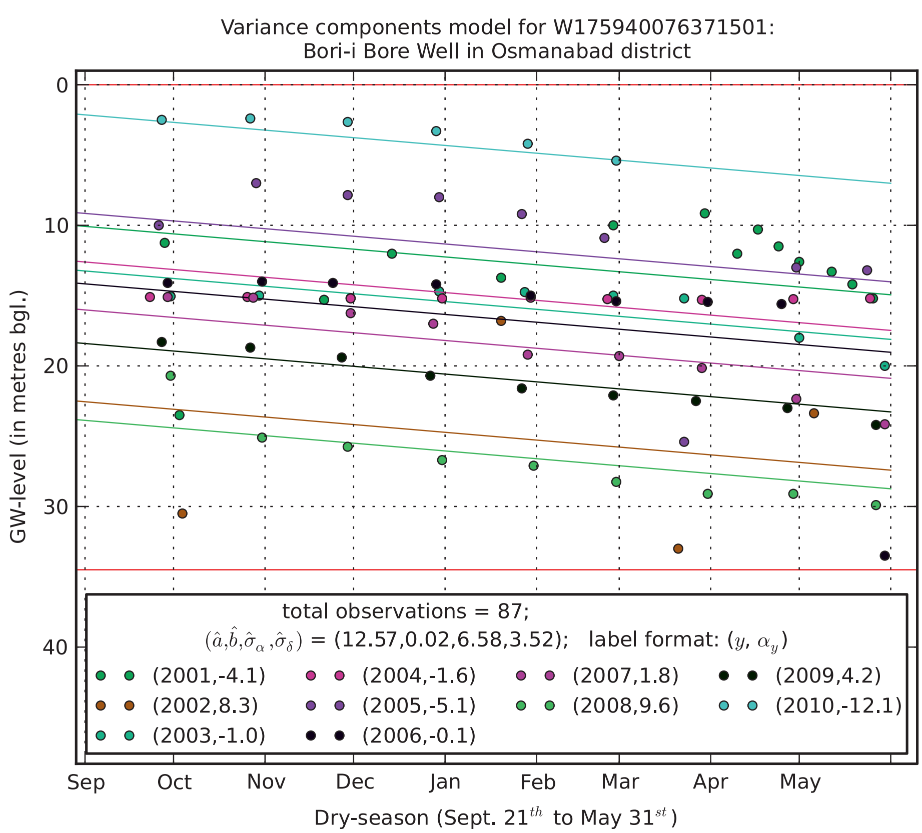

Figure 5. Example plot of variance components model. Estimated offset trend for year y: $\hat{l}(y,t) = \hat{a} + \hat{b}t + \hat{\alpha}_y$; estimated residual for individual observations: $\hat{\delta}_{y,t}$ (not shown explicitly).

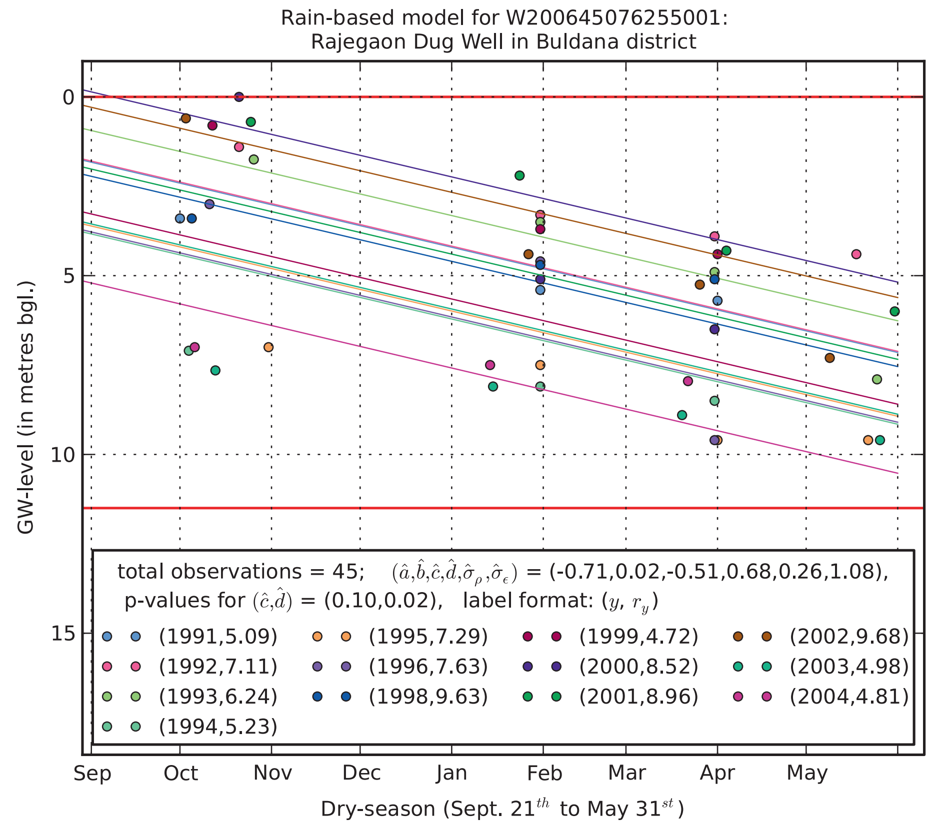

Figure 6. Typical graphical plot of rain-based model. Only predicted trend-lines (i.e., $\hat{l}(y,t) = \hat{a} + \hat{b}t + \hat{c}r_y + \hat{d}v_y$) are shown.

In our implementation, r is measured in units of 100 mm, while groundwater levels are measured in metres. So, the quantities c, ρ and σ_ρ, which refer to the rainfall-effect, have units 'metres of groundwater level per 100 mm of rainfall'. For brevity, this unit is implied, whenever the values of c, ρ and σ_ρ are mentioned unqualified.

5.3.1 *Results about factor-effect parameters c and d*

The average effect of the rainfall parameter r is found statistically significant (at 10%) for 1879 sites. It is worth noting that, although monsoon rainfall is the primary source of groundwater-recharge, r does not significantly determine the aggregate dry-season groundwater-level for the remaining wells, which form a majority. This may happen when the yearly-effect of rainfall varies excessively (σ_ρ is high) compared to the average effect (c) or when the net groundwater-regime is largely insensitive to variation in total monsoon rainfall (for example, in the western coastal region where σ_α is close to zero). In any case, this shows that it may not always be appropriate to rely on the amount of monsoon rainfall to predict dry-season groundwater-availability.

The average effect is positive (dry-season groundwater levels higher for higher value of r) for 1781 of the 1879 sites having significant effect, while it is negative for the remaining 98 sites. The latter may form interesting case-studies for the possibility of how the phenomenon of variation in aggregate dry-season groundwater resource, which is generally believed to be dominantly hydrogeological, is significantly affected by community behaviour. Figure 4 also contains a rasterized map generated using values of $-\hat{c}$ for all modelled wells. Note that $c < 0$ is positive effect since groundwater levels are measured below ground-level; hence map of $-\hat{c}$.

The effect of pre-monsoon groundwater level v is found statistically significant (at 10%) for 1806 sites. The effect is positive (dry-season groundwater levels higher for higher pre-monsoon groundwater levels) for 1689 of these sites, while for the remaining 117 it is negative. In figure 4, the rasterized map obtained using the values of $\hat{d}$ at the modelled sites, does not show macro-scale regional patterns as seen in the earlier maps.

5.3.2 *Results about variance parameters $\hat{\sigma}_\rho$ and $\hat{\sigma}_\epsilon$*

A total of 546 sites had a negative estimate of σ_ρ^2, indicating that actual variance value must be close to zero. Figure 4 includes a rasterized map of spatial distribution of $\hat{\sigma}_\rho$ in Maharashtra. Both, $\hat{\sigma}_\epsilon$ and $\hat{\sigma}_\delta$, are essentially the estimates of intraseasonal variance. So, their results are similar as seen from

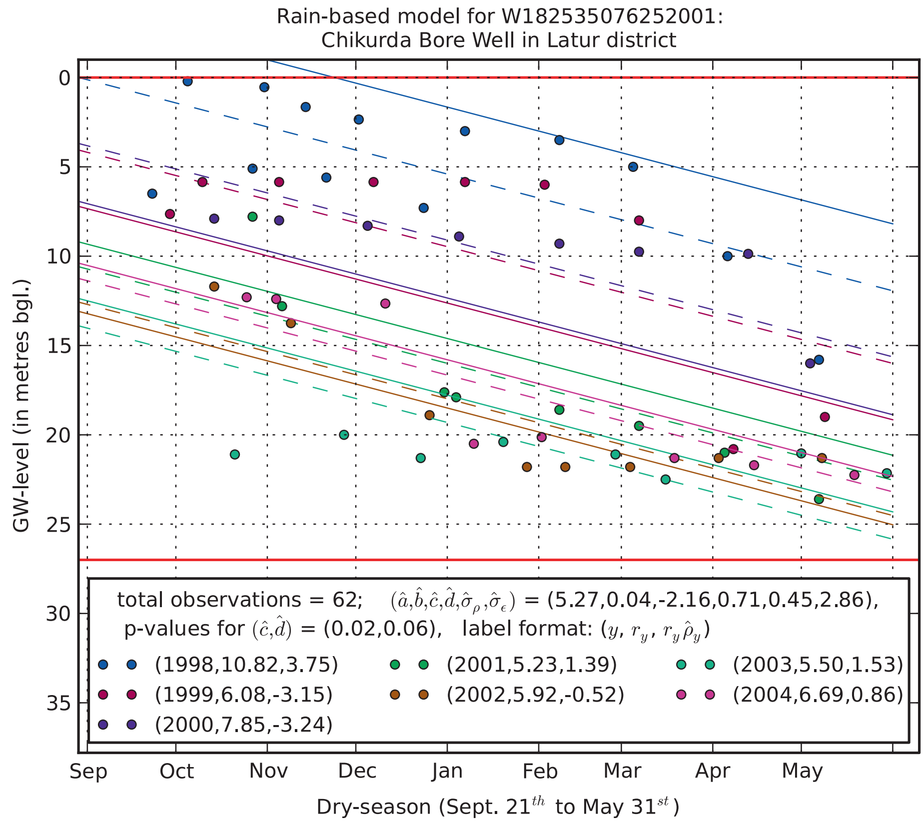

Figure 7. Typical graphical plot of rain-based model. Both predicted trend-line (bold) as well as trend-line with estimated residual offset (dotted) (i.e., $\hat{l}(y,t) = \hat{a} + \hat{b}t + (\hat{c} + \hat{\rho}_y)r_y + \hat{d}v_y$) are shown.

figure 4. Also, specifically, no site had a negative estimate of σ_ϵ, while 14 sites had $\sigma_\epsilon > 10$ m.

5.3.3 *The residual offsets $r_y\rho_y$*

The unexplained yearly variation is accounted into the residual $r_y\rho_y$. Akin to the statistic of root mean squared error (RMSE), we may measure the average magnitude of these residual offsets using a root mean squared residual offset statistic defined as:

$$\text{RMSRO} = \sqrt{\frac{\sum_{y\in Y}(r_y\hat{\rho}_y)^2}{|Y|}}$$

where Y is the set of years of observations. It is desirable that RMSRO be small so that aggregate dry-season groundwater levels are sufficiently determined by r and/or v.

Figure 4 includes a rasterized map for the spatial distribution of RMSRO. As per our model, this is the groundwater level variance that remains unexplained by aggregate rainfall or previous year's level, and thus may be attributed to other variations in groundwater-regime conditions that occur across the years.

Modeling results for all three models of each observation well are available at www.gise.cse.iitb.ac.in/gsda.

5.4 *Regional characterization of groundwater-level variation*

The conspicuous resemblance among the spatial patterns of $\hat{\sigma}_\alpha$, $\hat{\sigma}_\delta$, $-\hat{c}$, $\hat{\sigma}_\rho$, RMSRO, in figure 4 leads us to a zoning of Maharashtra, as shown in figure 8. The characteristics of our model parameters as well as an identifying label for the four zones are provided in table 2.

Note that these patterns are closely related to physical features of the four zones. For example, the coastal belt (Zone 1) receives heavy monsoon rainfall of more than 1500 mm, but has low aquifer depths of about 5–15 m (see GSDA 2010). The initial spells of monsoon rainfall saturate these aquifers. Further variations in r have no effect $(c \sim 0)$ and σ_α remains negligible. Zone 2 receives low rainfall and is under the Drought Prone Area Programme (DPAP) of the Govt. of India. A possible sequence of events leading to high variance may begin with the watershed interventions created in the DPA programme to improve recharge. The subsequent improvement in groundwater from such measures, which may not be accurately estimated, and an unregulated cropping pattern, may lead to high values of σ_ρ and hence to recharge and discharge cycles of high variance. In Zone 3, which has better groundwater capacity and fertile

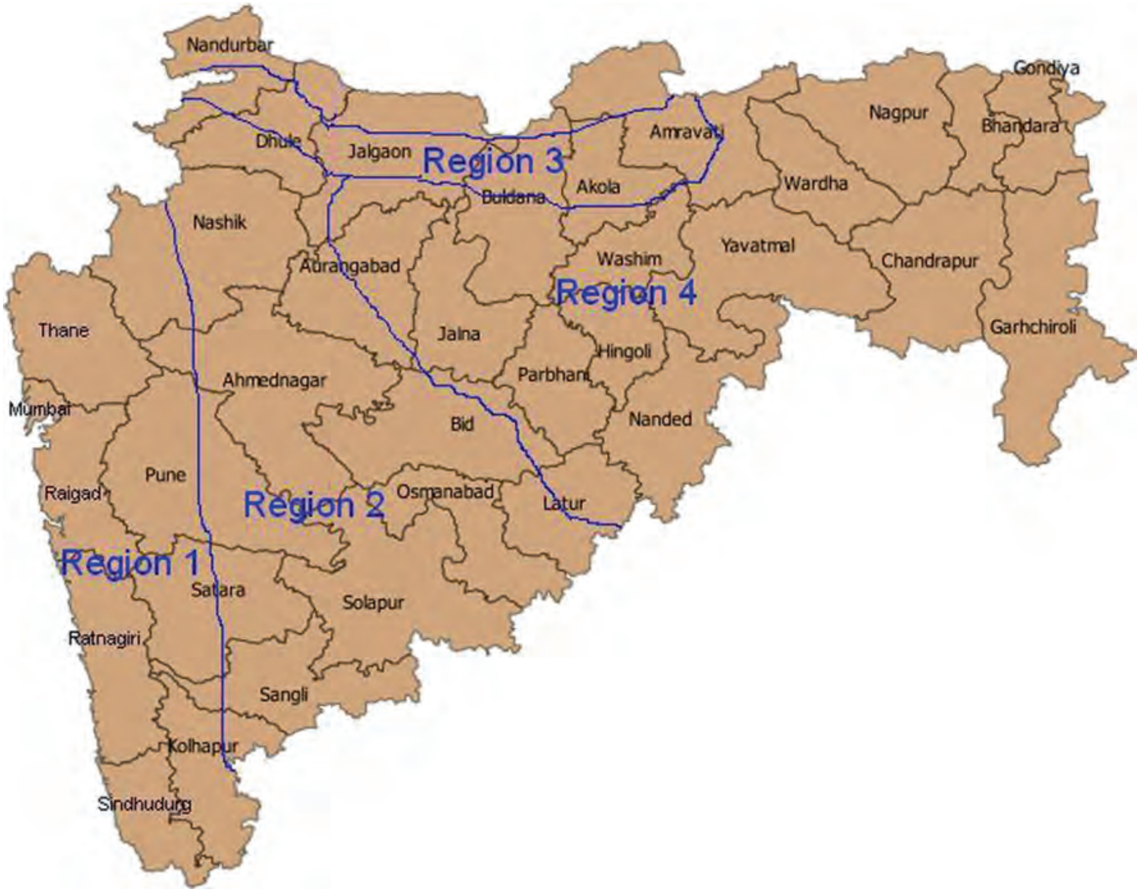

Figure 8. Zoning of Maharashtra based on distinct characteristics of groundwater-level variation in the recent decades.

Table 2. *Broad summary of the model parameters for the four zones.*

	Zone 1 (Coastal districts and Western Ghats)	Zone 2 (DPAP zone)	Zone 3 (Tapi–Purna alluvial deposition belt)	Zone 4 (Rest of Maharashtra excluding zones 1, 2 and 3)
$\hat{\sigma}_\alpha$	Neglegible	Very high	Very high	Moderate to low
$\hat{\sigma}_\delta$ (similar to $\hat{\sigma}_\epsilon$)	Very low	High	High	Low
$-\hat{c}$	Very low	High	Moderate	Low
$\hat{d}$	Generally low	No pattern	No pattern	No pattern
$\hat{\sigma}_\rho$	Very low	Extremely high	Very high	Low to very low
RMSRO	Very low	High	Moderately high	Low to very low

alluvium of the Tapi–Purna basin (see GSDA 2010), the availability of affordable drilling and pumping technology may have triggered similar unregulated and competitive extraction, thus leading to high variance.

A snapshot of forest cover in the year 2009 is shown in figure 9, which is also an indication of low anthropogenic activity. Resemblence between forest cover and low variances is seen in the eastern part of Zone 4 and northern part of Amravati district and parts of Zone 1.

6. Discussion

Some important notes about our models are jotted down in subsection 6.1. Subsection 6.2 briefly highlights the importance of our findings for the administration of groundwater.

6.1 *Notes on modelling*

- The assumption of normal distribution is frequently used when the statistical nature of the

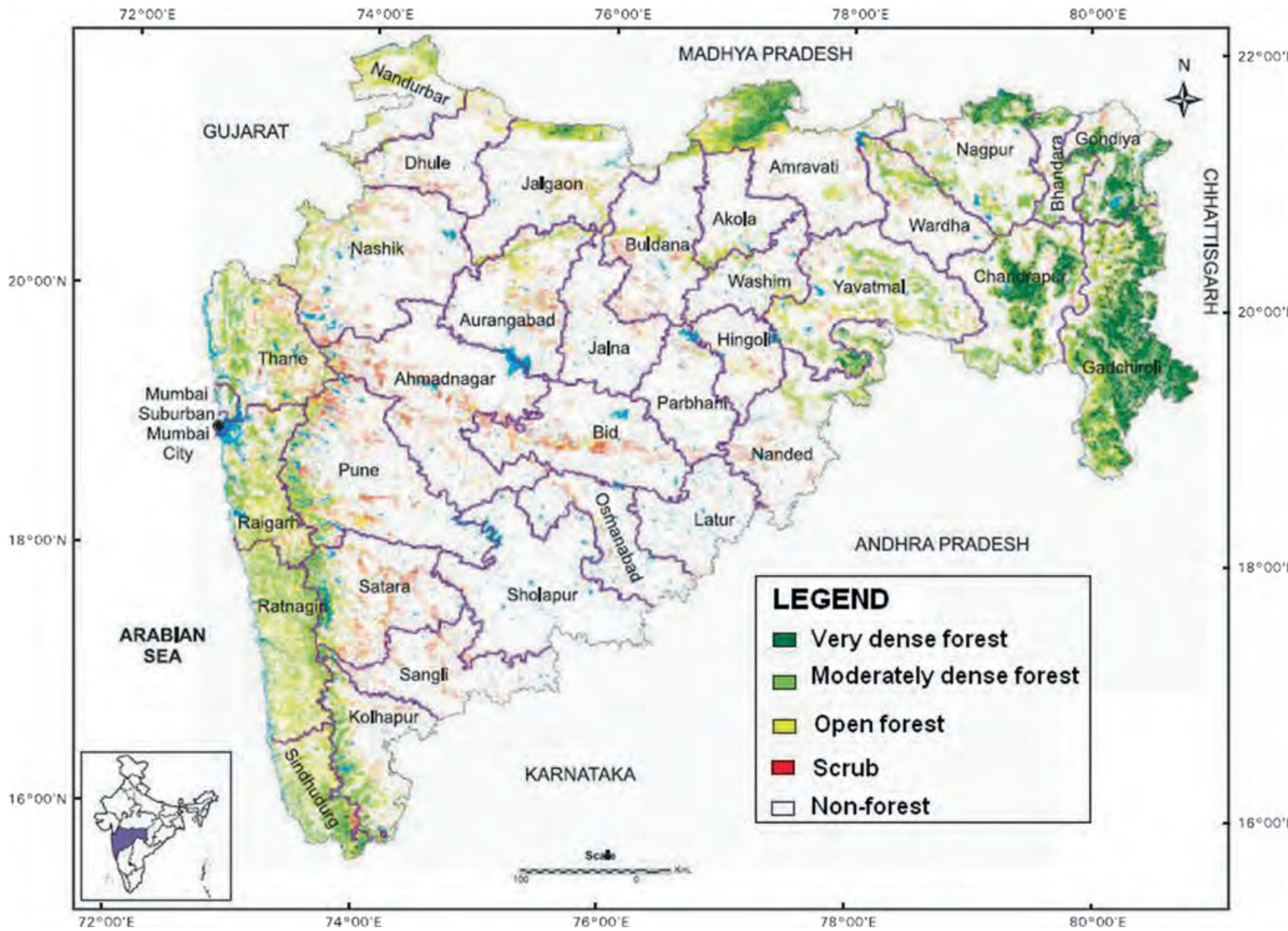

Figure 9. Forest cover map of Maharashtra as per year 2009. Reproduced from FSI (2009).

modelled phenomenon is not clearly known. The Kolmogorov–Smirnov test for deviation from normality has been applied at 10% significance, for sites that have at least 10 samples (years for $\hat{\alpha}_y$, $\hat{\rho}_y$ and observations for $\hat{\delta}_{y,t}$, $\hat{\epsilon}_{y,t}$) available. From the results presented in table 3, it seems reasonable to consider normal distribution (at least for interannual variances, which is the focus of this paper) for general state-wide assessment.

- Variance components model usually refers to random-effects models. Our so-called variance components model is actually a mixed-effects model, with a and b as the fixed-effects and α_ys as the random-effects. The name used here, is only meant to signify its purpose of classifying groundwater-level variance (unexplained by the obvious $a + bt$).
- Parameter estimates of (conceptually) common model parameters like b, intraseasonal variance, etc., need not be same in the three models, although they are usually close. This happens because multilevel model estimation, implicitly and rightly, assigns more importance to those groups (years, in our case), which have more number of observations (see Gelman and Hill 2007, for detailed description). Furthermore, in random coefficient modelling, the coefficient 'errors' (variations ρ_ys, in our case) in the model formulation are nonuniformly weighted due to the yearly multipliers (r_y, in our case). Thus, although the three models form a sequence of conceptual refinements, they do not formally fit into a single framework of mathematical refinements, as is the case with multiple regression models with increasing number of regressors. It is not clear if such a framework exists for multilevel regression and is computable. All the same, this has little impact upon our inferences.

Table 3. *Results of KS-test for the estimated values of random deviations.*

Significant (at 10%) deviation from normality for	No. of wells rejected	No. of wells accepted
$\hat{\alpha}_y$	4085	402
$\hat{\delta}_{y,t}$	4021	1232
$\hat{\rho}_y$	2882	372
$\hat{\epsilon}_{y,t}$	4061	601

- The effect of monsoon rainfall depends not only on the total amount, but also on the pattern of daily rainfall. We have used total monsoon rainfall r since it is the most natural and widely accepted numeraire and is also used by the state administration. Our model framework is easily adaptable for use of any other numeraire w, by simple replacement of r by w. However, this would be more useful if rainfall was measured at observation wells, which is currently not done.
- Ideally, the fluctuation of model's groundwater levels should be bounded by the ground-surface and aquifer depth. In the absence of such bounding, if actual groundwater levels were repeatedly hitting these limits and affecting model estimation, this would appear as a correlation between the regressor (e.g., r_y) and the estimate of corresponding error ($\hat{\rho}_y$ for r_y). However, significant correlations between r_y and $\hat{\rho}_y$ were practically absent in almost all analyzed wells, and the inference about lack of significance of rainfall factor is not affected.

The last two items point to some current technical limitations of our models.

6.2 *Scarcity and uncertainty*

It is necessary to recognize that scarcity and uncertainty are mutually distinct features of a groundwater regime. For example, if groundwater was scarce but certain, the groundwater-user may make a different set of socio-economic decisions as compared to when it were both scarce and uncertain. In the first case, it incentivizes efficient use of groundwater, while in the second case, it may well lead to competitive extraction and a race to the bottom, worsening the scarcity. Indeed, the spatial coincidence of large σ_α with σ_ρ seems to suggest this.

This leads us to the following policy recommendations: (i) recognition of scarcity and uncertainty as separate attributes of groundwater-availability and developing indices to measure uncertainty, (ii) further work into the incorporation of socio-economic data along with hydrogeologic and climatic data for building groundwater assessment tools. Perhaps, one relevant avenue for this is the periodic water balance computation carried out by GSDA for each watershed, every 3–5 years (see GEC'97 1997). This incorporates considerable data on extraction, irrigation, surface water bodies, and estimates of other stocks and flows. A refinement of this water balance exercise may yield better inputs for the yearly outlook for the dry season.

7. Conclusions

Data gathered by government agencies, at different scales and for different administrative or geographical units, and with inherent weaknesses, is used almost universally to inform policy. This makes the analysis of this data and strengthening the data gathering processes an important scientific objective. This study of the GSDA's observation well dataset is a step in that direction.

Multilevel statistical regression is aptly used in our empirical analysis. It reveals that the typical groundwater-user faces not only scarcity but also uncertainty in the access to such a key resource. The study classifies uncertainty as that arising within the year and that arising across years, and shows that rainfall is a poor predictor of groundwater levels. This points to a complex relationship between groundwater users, their perception of hydroclimatic conditions and the decisions that they make on land-use and cropping, across years. This needs to be better understood if a scientific groundwater regime is to be implemented. One of the limitations of this work is the use of the IMD aggregate rainfall dataset as a numeraire, for want of better data. Perhaps, a future study of a limited number of observation wells, where daily rainfall data is actually available, may sharpen this understanding.

Another outcome of the study is the construction of key parameters (b, c, ρ, d) which represent the aggregate balance of hydrogeologic and anthropogenic phenomena and are a measure of the health of the local groundwater regime. These parameters may serve to set targets for mitigation and policy interventions. Calculation of these parameters in a set of representative and more carefully observed regimes may help in their calibration and adoption as policy guidelines. Thus, our work points out that a more integrated socio-economic, hydrogeological and climatological data gathering and its analysis would lead to better models and sharper recommendations.

Acknowledgements

The groundwater-level data for the observation wells in Maharashtra was generously provided by GSDA. Thanks also to Prof. Subimal Ghosh (Dept. of Civil Engg., IITB, Mumbai, India), who provided IMD's daily rainfall dataset from his research repository.

References

Adyalkar P G and Mani V V S 1971 Paleogeography, geomorphological setting and groundwater possibilities in the Deccan Traps of western Maharashtra; *Bull. Volcanologique* **35(3)** 696–708.

Banerji A, Meenakshi J V and Khanna Gauri 2012 Social contracts, markets and efficiency: Groundwater irrigation in North India; *J. Dev. Econ.* **98(2)** 228–237.

Bose R N and Ramakrishna T S 1978 Electrical-resistivity surveys for groundwater in the deccan trap country of Sangli District, Maharashtra; *J. Hydrol.* **38** 209–221.

Demissie Yonas Kassa 2008 Data-driven models to enhance physically-based groundwater model predictions; Doctoral Dissertation, University of Illinois at Urbana-Champaigne.

Deolankar S B 1980 The Deccan Basalts of Maharashtra, India – Their potential as aquifers; *Ground Water* **18(5)** 434–437.

Forest Survey of India, Govt. of India 2009 India State of Forest Report 2009.

GEC (Groundwater Estimation Committee) 1997 Groundwater Estimation Methodology – 1997, Ministry of Water Resources, Govt. of India.

Gelman Andrew and Hill Jennifer 2007 Data Analysis Using Regression and Multilevel/Hierarchical Models; Cambridge University Press.

Giustolisi O and Savic D A 2009 Advances in data-driven analyses and modeling using EPR-MOGA; *J. Hydroinformatics* **11(3-4)** 225–236.

Goldstein H 1995 Multilevel Statistical Models, 2nd edn, London, Institute of Education, http://www.bristol.ac.uk/cmm/team/hg/msm-2nd-ed/.

Gore K P, Pendke M S, Gurunadha Rao V V S and Gupta C P 1998 Groundwater modeling to quantify the effect of water harvesting structures in Wagarwadi watershed, Parbhani District, Maharashtra, India; *Hydrol. Process.* **12** 1043–1052.

GSDA (Groundwater Surveys and Developmet Agency) of Govt. of Maharashtra & Central Groundwater Board (Central Region) of Govt. of India 2010 Report on the dynamic groundwater resources of Maharashtra (2008–2009).

GSDA (Groundwater Surveys and Development Agency)'s Kolhapur district unit, 2013 Plan for the possible water deficit for year 2013–2014 (dry-season) based on rainfall and groundwater levels; http://www.cse.iitb.ac.in/~rahulbg/dryseason_predicted_water_deficit.pdf.

Kaplay R D and Patode H S 2004 Groundwater pollution due to industrial effluent at Tuppa, New Nanded, Maharashtra, India; *Environ. Geol.* **46** 871–882.

Kholghi M and Hosseini S M 2009 Comparison of groundwater level estimation using neuro-fuzzy and ordinary kriging; *Environmental Modeling & Assessment* **14(6)** 729–737.

Kulkarni Himanshu, Deolankar S B, Lalwani Anil, Joseph Bijoy and Pawar Suresh 2000 Hydrogeological framework of the Deccan basalt groundwater systems, west-central India; *Hydrogeol. J.* **8** 368–378.

Kulkarni Himanshu, Shankar P S Vijay, Deolankar S B and Shah Mihir 2004 Groundwater demand management at local scale in rural areas of India: A strategy to ensure water well sustainability based on aquifer diffusivity and community participation; *Hydrogeol. J.* **12** 184–196.

Lallahema S, Maniaa J, Hania A and Najjarb Y 2005 On the use of neural networks to evaluate groundwater levels in fractured media; *J. Hydrol.* **307(1–4)** 92–111.

Limaye Shrikant Daji 2010 Review: Groundwater development and management in the Deccan Traps (basalts) of western India; *Hydrogeol. J.* **18** 543–558.

Naik Pradeep K and Awasthi Arun K 2003 Groundwater resources assessment of the Koyna river basin, India; *Hydrogeol. J.* **11** 582–594.

Passarella Giuseppe, Barca Emanuele and Vurro Michele 2013 Optimal redesign of groundwater monitoring networks: A data-driven multi-approach software, In: Conference proceedings of the 8th International Conference of EWRA: Water Resources Management in an Interdisciplinary and Changing Context, Porto, Portugal.

Pawar N J, Pondhe G M and Patil S F 1998 Groundwater pollution due to sugar-mill effluent, at Sonai, Maharashtra, India; *Environ. Geol.* **34(2–3)** 151–158.

Phansalkar Sanjiv and Kher Vivek 2006 A decade of the Maharashtra groundwater legislation: Analysis of the implementation process, 2/1 Law; *Environment and Development J.*, p. 67, http://www.lead-journal.org/content/06067.pdf.

Rajeevan M and Bhate J 2009 A high resolution daily gridded rainfall dataset (1971–2005) for meso-scale meteorological studies; *Curr. Sci.* **96(4)** 558–562.

Rokade V M, Kundal P and Joshi A K 2007 Groundwater potential modelling through remote sensing and GIS: A case study from Rajura Taluka, Chandrapur District, Maharashtra; *J. Geol. Soc. India* **69(5)** 943–948.

Sekhri Sheetal 2012 Sustaining groundwater: Role of policy reforms in promoting conservation in India; *India Policy Forum* 2012–13 **(9)** 149–187.

Sekhri Sheetal 2014 Wells, water, and welfare: Impact of access to groundwater on rural poverty and conflict; *Appl. Econ.* **6(3)** 76–102.

Shepard D 1968 A two-dimensional interpolation function for irregularly-spaced data; Proceedings of the 1968 ACM National Conference, pp. 517–524.

Shirmohammadi Bagher, Vafakhah Mehdi, Moosavi Vahid and Moghaddamnia Alireza 2013 Application of several data-driven techniques for predicting groundwater level; *Water Resour. Manag.* **27** 419–432.

MS received 18 September 2014; revised 10 January 2015; accepted 12 January 2015