

# JGR Solid Earth

## RESEARCH ARTICLE

10.1029/2021JB022373

### Special Section:

Machine learning for Solid Earth observation, modeling and understanding

### Key Points:

- Discontinuities in the long-term, seasonal amplitude and phase of geodetically measured displacement fields coincide with fault strands
- The belt of high strain rate is characterized by the low-to-moderate vegetation fraction
- Phase discontinuities in seasonal displacement occur near the transition between sedimentary basins and consolidated ranges

### Supporting Information:

Supporting Information may be found in the online version of this article.

### Correspondence to:






X. Hu,  
[hu.xie@pku.edu.cn](mailto:hu.xie@pku.edu.cn)

### Citation:

Hu, X., Bürgmann, R., Xu, X., Fielding, E., & Liu, Z. (2021). Machine-learning characterization of tectonic, hydrological and anthropogenic sources of active ground deformation in California. *Journal of Geophysical Research: Solid Earth*, 126, e2021JB022373. <https://doi.org/10.1029/2021JB022373>

Received 8 MAY 2021  
Accepted 27 OCT 2021

## Machine-Learning Characterization of Tectonic, Hydrological and Anthropogenic Sources of Active Ground Deformation in California

Xie Hu<sup>1,2,3</sup> , Roland Bürgmann<sup>1</sup> , Xiaohua Xu<sup>4</sup> , Eric Fielding<sup>5</sup> , and Zhen Liu<sup>5</sup> 

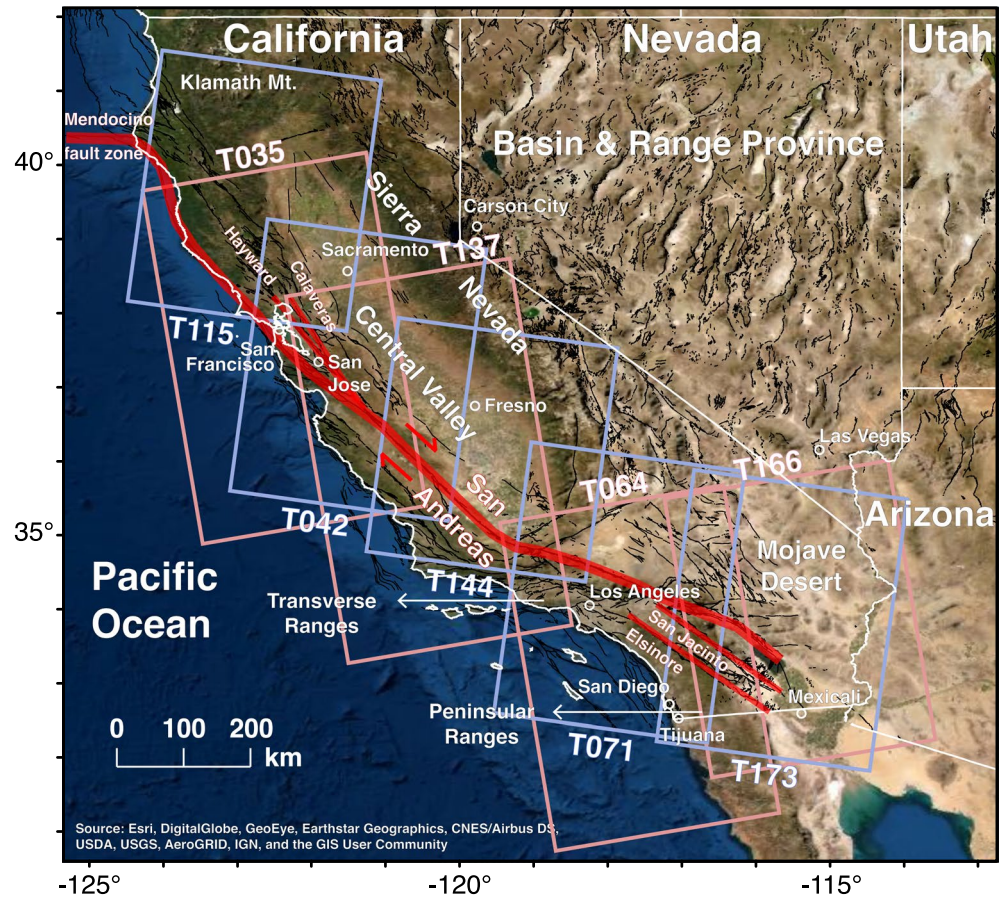
<sup>1</sup>Berkeley Seismological Laboratory and Department of Earth and Planetary Science, University of California, Berkeley, CA, USA, <sup>2</sup>Department of Civil and Environmental Engineering, University of Houston, Houston, TX, USA, <sup>3</sup>Now at the College of Urban and Environmental Sciences, Peking University, Beijing, China, <sup>4</sup>Institute for Geophysics, University of Texas at Austin, Austin, TX, USA, <sup>5</sup>Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

**Abstract** Tectonic, hydrological and industrial processes coexist in the dynamic natural environments. However, our knowledge of ground deformation associated with tectonic, hydrological and anthropogenic processes and their interactions remains limited. California represents a natural laboratory that hosts the San Andreas fault system, Central Valley and other aquifer systems, and extensive human extraction of natural resources. The attendant multi-scale ground deformation that has been mapped using Copernicus Sentinel-1 Synthetic Aperture Radar (SAR)-satellite constellation from four ascending and five descending tracks during 2015–2019. We consider the secular horizontal surface velocities and strain rates, constrained from GNSS measurements and tectonic models, as proxies for tectonic processes, and seasonal displacement amplitudes from interferometric SAR (InSAR) time series as proxies for hydrological processes. We synergize 23 types of multidisciplinary datasets, including ground deformation, sedimentary basins, precipitation, soil moisture, topography, and hydrocarbon production fields, using a machine learning algorithm—random forest, and we succeed in predicting 86%–95% of the representative data sets. High strain rates along the SAF system mainly occur in areas with a low-to-moderate vegetation fraction (~0.3), suggesting a correlation of rough/high-relief coastal range morphology and topography with the active faulting, seasonal and orographic rainfall, and vegetation growth. Linear discontinuities in the long-term, seasonal amplitude and phase of the surface displacement fields coincide with some fault strands, the boundary zone between the sediment-fill Central Valley and bedrock-dominated Sierra Nevada, and the margins of the inelastically deforming aquifer in the San Joaquin Valley, suggesting groundwater flow interruptions, contrasting elastic properties, and heterogeneous hydrological units.

**Plain Language Summary** Although scientific advances have been achieved in every individual geoscience discipline with more extensive and accurate observations and more robust models, our knowledge of the Earth complexity remains limited. The tectonic, hydrological and anthropogenic processes interact in the highly populous California, which have contributed to multi-scale ground deformation. Here we rely on remotely sensed ground deformation products and locations of oil and gas fields as proxies for tectonic, hydrological and anthropogenic processes. The training model of the random forest algorithm has a good performance in predicting the representative multidisciplinary datasets and reveals their intrinsic similarities and relative importance in the predictions. We also note compelling spatial correlation between the long-term and seasonal displacement discontinuities, the high-strain-rate fault zones, and a narrow range of vegetation fraction, as well as the margins of the heterogeneous structures.

## 1. Introduction

Earth's surface and interior processes, including tectonic faulting, subsurface water storage, compacting sediments, and human extraction and injection of fluids jointly modulate the Earth's deformation. In particular, California hosts the San Andreas fault (SAF) system that separates the Pacific Plate and the North American Plate and accommodates broadly distributed transform motions of ~45 mm/year (DeMets et al., 2010). Fault slip rates of the many fault strands that make up the SAF system have been constrained



**Figure 1.** California landscape and tectonic setting. Red and purple boxes show the footprint of ascending (T035, T137, T064, and T166) and descending (T115, T042, T144, T071, and T173) Sentinel-1 tracks, respectively. Black lines show the fault traces obtained from USGS/CGS. Red lines show the principal fault zones of the San Andreas fault system.

from dated offset geologic markers (Dawson & Weldon, 2013 and references cited therein) as well as mechanical modeling of geodetic measurements of interseismic surface deformation (Evans, 2018; Lindsey & Fialko, 2013; Lundgren et al., 2009; Molnar & Dayem, 2010; Tong et al., 2013, and references cited therein) (Figure 1). Dozens of M6.5+ earthquakes have struck California since 1800. The most recent surface-rupturing event is the M7.1 Ridgecrest earthquake that occurred ~200 km northeast of Los Angeles on July 5, 2019, preceded by a M6.4 foreshock one day prior (Ross et al., 2019; Xu et al., 2020).

During the late Pliocene and Quaternary, strike-slip and reverse fault systems gave rise to the mountain ranges surrounding coastal plains and inland basins, which host the present-day aquifer systems. The Central Valley is one of the principal aquifer systems in the U.S. that encompasses 52,000 km<sup>2</sup> across northern and central California (Faunt, 2009). Because of increasing urbanization, human extraction of natural resources, such as groundwater, hydrocarbons and geothermal fluids, have been increasing in the region. For example, groundwater overexploitation led to prolonged groundwater depletion and permanent compaction in the San Joaquin Basin starting in the 1920s (Faunt, 2009; Poland et al., 1975). A recovery plan in the early 1970s restored the groundwater storage in most aquifers; however, increased water pumping during drought periods of 1976–1977, 1987–1992, and more recent 2006–2010 and 2012–2016 reversed the groundwater-level recovery and aggravated the subsidence (e.g., Diffenbaugh et al., 2015; Faunt, 2009; Faunt et al., 2016; Liu et al., 2019; Smith et al., 2017).

Land subsidence can impact lifelines and waterways and increase the risks of flooding. Overexploitation makes some of the underground reservoirs permanently lose capacity to store groundwater (e.g., Ojha et al., 2018; Smith et al., 2017). The associated changes in pore pressure and mass unloading may modulate

the shallow deformation and stress fields in the crust and affect the deeper seismogenic zone (e.g., Carlson et al., 2020; Hu & Bürgmann, 2020; Johnson et al., 2017). In particular, the anthropogenic operations in hydrocarbon and geothermal fields have been shown to trigger earthquakes, including damaging events, as fluid is pumped both into and out of a reservoir (e.g., Brodsky & Lajoie, 2013; Goebel et al., 2015; Goebel & Shirzaei, 2020; Hough & Bilham, 2018; Hough & Page, 2016; Johnson et al., 2016; Trugman et al., 2016). California has a large area and population and complex nature of the tectonic, hydrological and anthropogenic strain sources. Systematic mapping and characterization of deformation associated with fault systems, groundwater basins and industrial fields in the midst of plate-boundary active tectonics are important. We aim to improve our understanding of the interconnection among natural and anthropogenic processes for enhanced groundwater and earthquake hazard assessment, and for the mitigation of associated socio-economic risks.

Human discovery and understanding of the Earth system are now being transformed by big data. Improved techniques on data processing and high-performance storage and computation have triggered a breakthrough in deciphering the code of nature (Bergen et al., 2019). Scientific advances have been achieved in most individual earth science disciplines with more accurate observations and more robust models. However, our knowledge of the atmosphere, hydrosphere, and lithosphere has not been systematically linked, although multidisciplinary data sets are readily available, and are especially complete and robust in California (Small et al., 2017; Zeng & Shen, 2017). Thus, California is an ideal natural laboratory for using big data in cross-disciplinary research.

As a subdivision of artificial intelligence, machine learning (ML) offers a promising solution to deal with the broad channels of nonlinear and nonparametric Earth signals in an empirical fashion. The fundamental idea is that the selection and decision-making criteria based on a considerable number of features as an entity will outperform any individual entity. The ML models can investigate the correlations of features that appear to provide a predictive capability. Classic ML algorithms include neural networks, support vector machines, self-organizing maps, decision trees and random forest (RF) among others (Lary et al., 2016). Although geoscientists have been using ML to address various geoscientific problems for decades (e.g., Dowla et al., 1990), it remains in the early stage of its development as many long-standing and unconventional data sources are largely unexplored and the considerable volume and diverse character of information remains fragmented and incomplete (Bergen et al., 2019). In this study, we rely on an ensemble learning algorithm by growing RF to combine remote sensing datasets from the subsurface, ground, and space, in an attempt to better understand the interconnection between the tectonics, lithology, hydrology, climate, and human activities in California.

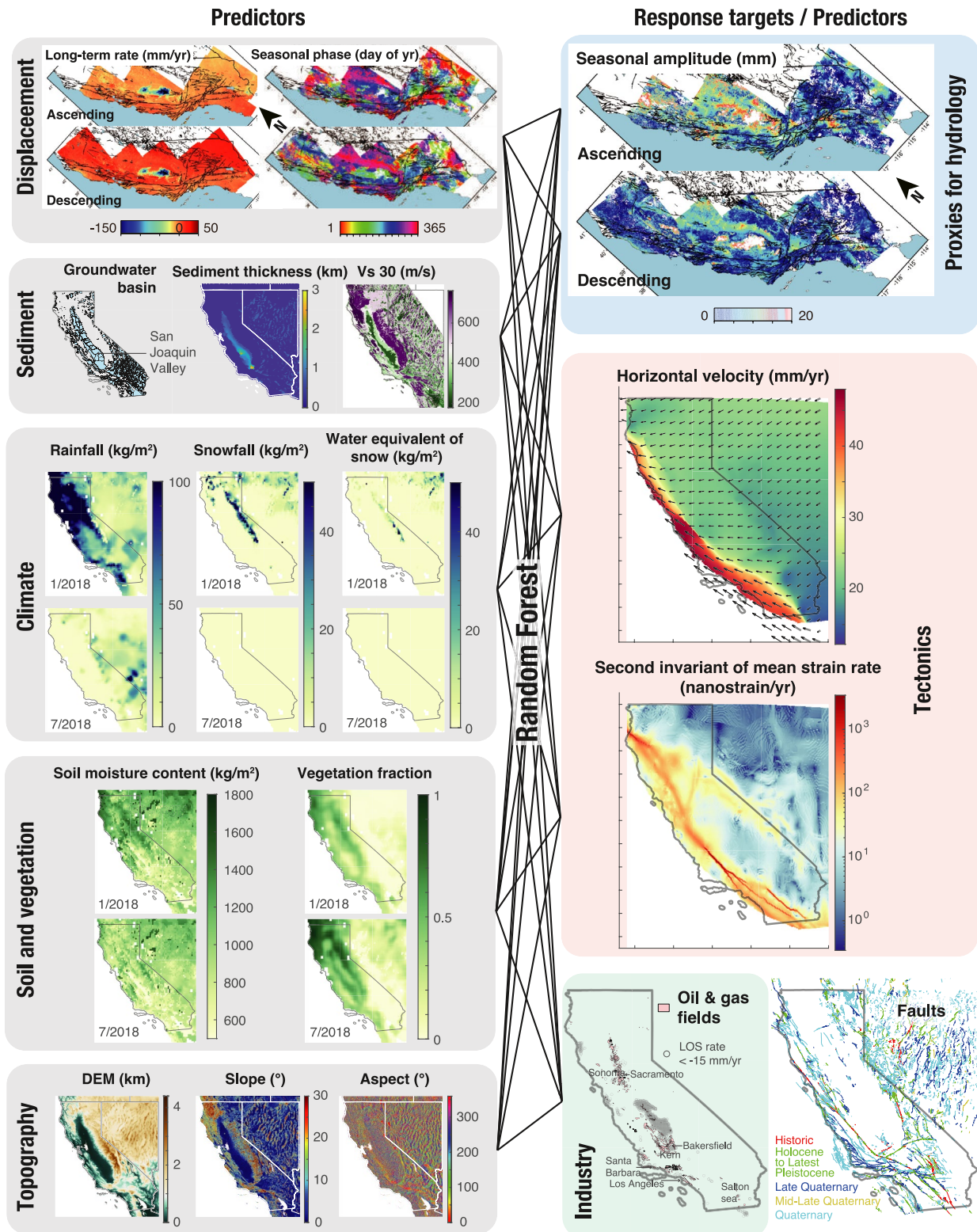
## 2. Data and Methods

### 2.1. Interdisciplinary Datasets

We collect various data sets as proxies for the tectonic, hydrological, and industrial processes (Figure 2; Table S1 in Supporting Information S1): the ascending and descending InSAR line-of-sight (LOS) rate, seasonal amplitude and phase; tectonics-related interseismic horizontal ground displacement rate and second invariant of strain rate derived from GPS measurements and geologic fault slip-rate estimates; sedimentary basin-related data including the geographic locations of basin-fill deposits, the sediment thickness, and shear wave velocity at a depth of 30 m ( $V_{s30}$ ); precipitation data including rainfall, snowfall and water equivalent thickness of snow; soil moisture and vegetation fraction; topographic features including elevation, slope, and aspect; and the geographic locations of oil & gas fields, fluid extraction and/or injection wells, and tectonic faults.

#### 2.1.1. Time-Series InSAR Products

The European Space Agency launched the Copernicus Sentinel-1A satellite in 2014, followed by the Sentinel-1B satellite in 2016. This C-band twin-satellite constellation has started a new era for time-series InSAR data processing and analysis with regular repeat cycle of 6–12 days from both ascending and descending orbits over California. Xu and Sandwell from SIO at UC San Diego produced 2015–2019 time-series ground displacement products from GMTSAR analysis of Sentinel-1 collection (<https://topex.ucsd.edu/gmtsar/in-sargen/>; Xu et al., 2021), including four ascending consecutive tracks (T035, T064, T137 and T166) and five



**Figure 2.** Schematic view of the growing random forest using multidisciplinary datasets. See Section 2.1 and Table S1 in Supporting Information S1 for details on each dataset and its source.

consecutive descending tracks (T042, T071, T115, T144 and T173) (The red and purple boxes in Figure 1). Here we focus on the interseismic data collected before the 2019 Ridgecrest earthquake sequence, thus avoiding the subsequent co- and post-seismic deformation transients currently affecting much of southern California. The mapped displacements are measured along the radar line-of-sight (LOS), which is effectively insensitive to the north-south component due to near-polar orbits. Building on a considerable body of previous work, we use these products and focus on characterizing the long-term rate, as well as the seasonal peak-to-peak amplitude and phase of surface deformation.

### 2.1.2. Tectonic Data

A collection of five fault and fold layers of Historic, Holocene to Latest Pleistocene, Late Quaternary, Mid-Late Quaternary, and Quaternary is archived by the U. S. Geological Survey (USGS) and California Geological Survey (CGS) (<https://www.usgs.gov/natural-hazards/earthquake-hazards/faults>) (Figure 1 and Figure S1 in Supporting Information S1). The Quaternary time spans <2.58 Myr. The Holocene to Latest Pleistocene (10Kyr B.P to 1.65Myr B.P.) faults include a number of structures in the Mojave Desert in southern California.

Zeng and Shen (2017) published the crustal deformation field in the Western U.S. based on the joint inversion of horizontal GPS velocities and tectonic slip-rate constraints. We apply this horizontal deformation field over California and compute the second invariant of the strain rate (Kreemer & Hammond, 2007; Shen et al., 2015). We attribute this large-scale, horizontal strain-rate map to a pure tectonic origin, as the contemporary tectonic process is governed by the right-lateral strike-slip San Andreas fault system (See the second invariant of the mean strain rate in Figure 2). The hydrologically driven deformation processes, on the other hand, involve both elastic load deformation across the region and poroelastic deformation in basins. Both processes are predominately vertical (Hammond et al., 2016), but there are appreciable horizontal components in both near the margins of the hydrological loads and aquifer systems (e.g., Kreemer & Zaliapin, 2018; Shen & Liu, 2020).

### 2.1.3. Sedimentary Basin Data

We obtained the coverage of 515 groundwater basins and sub-basins compiled by the California Department of Water Resource ([http://atlas-dwr.opendata.arcgis.com/datasets/b5325164abf94d5cbeb48bb542fa616e\\_0](http://atlas-dwr.opendata.arcgis.com/datasets/b5325164abf94d5cbeb48bb542fa616e_0)) (Miller, 2000) (Figure 1 and Figure S2 in Supporting Information S1). The groundwater basins are filled with alluvial or unconsolidated deposits. California's primary aquifer system, the Central Valley, is comprised of the northern Sacramento Basin and the southern San Joaquin Basin (Faunt, 2009).

As a component of the U.S. Geological Survey National Crustal Model, a map of unconsolidated sediment thickness was generated from previous studies or derived directly from gravity analysis (<https://www.sciencebase.gov/catalog/item/5b0d85d4e4b0c39c934b0420>) (Shah & Boyd, 2018). We consider the sediment thickness as the first-order approximation of the thickness of aquifer-system deposits which are heterogeneous mixtures of unconsolidated to semi-consolidated gravel, sand, silt, and clay (Faunt, 2009).

Seismic shear-wave velocity ( $V_s$ ) at shallow depths is another indicator for rock and sediment type (e.g., Hsu et al., 2020). The average  $V_s$  of the top 30 m of strata ( $V_{s30}$ ) is an important criterion to help infer the stiffness of the materials and the ground amplification from earth shaking in the near surface (McPhillips et al., 2020). We retrieve  $V_{s30}$  in California from the CyberShake Study 17.3 Central California Velocity Model integrated with the Harvard Santa Maria and San Joaquin Basin Models in the Unified Community Velocity Model (UCVM) (<https://github.com/SCECcode/UCVMC>) (Small et al., 2017) from the Southern California Earthquake Center (SCEC). The Klamath Mountains, Coastal Ranges, Sierra Nevada, Transverse Ranges, and Peninsular Ranges from the north to the south present high velocities, while the Central Valley shows low velocities (Figure 2 and Figure S2 in Supporting Information S1). Alternating high and low velocities extending from the Sierra Nevada to Utah in the east, also present in some parts of the Mojave Desert, reflect the Basin and Range horst and graben structures and associated landscape. Low shear velocity regions are vulnerable to strong shaking and increased damage in the event of big earthquakes. As expected, the distribution of the principal groundwater basins, thick unconsolidated sediments, and low  $V_{s30}$  strongly overlap in the Central Valley and also smaller basins.

#### 2.1.4. Climatic and Soil-Moisture Related Data

Terrestrial hydrological systems are characterized by climatic and soil moisture-related features. We choose the North American Land Data Assimilation System-2 (NLDAS-2) Noah model data for monthly accumulated rainfall and snowfall, and monthly averages of shallow soil moisture and vegetation fraction (Figure 1, Figures S3 and S4 in Supporting Information S1). The spatial resolution is coarse with an interval of  $0.125^\circ$  (WGS84 datum), that is,  $\sim 14$  km.

Most of California has a Mediterranean climate. The temperature and precipitation depend largely on the latitude, morphology, elevation, and the proximity to the Pacific Ocean. The soil and climatic data exhibit strong seasonal variation. We consider January 2018 as a representative wet month and July 2018 as a representative dry month as demonstrated by the precipitation statistics (Figure 2, Figures S3, and S4 in Supporting Information S1).

Northern California receives heavier rain than the southern part of the state. High-altitude mountainous areas can be snow covered during much of the wet seasons. Ample seasonal rainfall nurtures the forests and grass- and brush-covered areas in the northern California exhibiting high seasonal vegetation fractions. The arid deserts, such as the Mojave Desert in southern California, are nearly void of vegetation. Soil moisture follows the distribution of rainfall, but the seasonal changes of the latter are more distinct. Soil and vegetation can maintain the moisture even without receiving precipitation for months. The vegetation in the Central Valley corresponds to the agricultural fields. Water recharge and discharge from anthropogenic activities (e.g., pumpage and irrigation) and natural hydrosphere modify the soil moisture content in space and time.

#### 2.1.5. Topographic Data Sets

We analyze the 3-arc-second (90-m resolution) Shuttle Radar Topography Mission (SRTM) version 3 digital elevation model (DEM) acquired in 2000 by NASA/JPL to characterize the topographic features, for example, elevation, slope angle, and slope aspect computed at the original 90-m resolution and then resampled to  $0.0025^\circ$  ( $\sim 270$  m) grids to be consistent with other features (Figure 1 and Figure S5 in Supporting Information S1), which characterize the locations of the basins and the ranges and the orientation of strike-slip faulting. The total elevation difference amounts to more than 4 km in California. Slope aspects represent the local directions of the largest topographic gradient, which have a salt-and-pepper appearance in the flat areas due to indistinct topographic relief and noise in the SRTM data.

#### 2.1.6. Oil and Gas Fields

Hydrocarbon and geothermal energy play an indispensable role in California's economy since the late nineteenth century. Here we only have and apply the geographic locations of hydrocarbon oil and gas fields from the Geologic Energy Management Division of the California Department of Conservation (CalGEM) (<https://gis-california.opendata.arcgis.com/datasets/cadoc::oil-and-gas-field-administrative-boundaries-1?geometry=-135.576%2C34.196%2C-103.936%2C40.311>). There are 240,574 active or inactive wells recorded in CalGEM (Figure 2 and Figure S6 in Supporting Information S1). The San Joaquin Valley contains the largest oil fields in California. Kern County and the Los Angeles Basin also have big concentrations of oil and gas production.

### 2.2. Spatiotemporal Features of Displacement Sources

As an imaging remote sensing approach with regular repeat cycles from orbiting satellites, InSAR can capture deformation features over a wide range of length and timescales. The spatiotemporal characteristics and three-dimensional patterns of the tectonic, hydrological and anthropogenic sources of ground motions differ systematically (Table 1).

Spatially, the tectonic, hydrological and anthropogenic signals decrease in size from thousands of  $\text{km}^2$  to hundreds of  $\text{m}^2$ . Tectonic strain accumulation along the San Andreas Fault system spreads out on the plate boundary scale. Interseismic creep may distribute along the fault trace or across a limited width around the fault while the elastic surface deformation off locked faults usually spreads out over tens of kilometers in a long-wavelength pattern. Lineations of high displacement gradients can be used to map and model the

**Table 1**

*Characteristics of Primary Displacement Sources in California*

Sources	Spatial features	Temporal features	Orientation
Tectonic	Plate-boundary scale with comparatively localized strain along active faults	Long-term rates at submillimeter to tens of millimeters per year on the interseismic timescales in the earthquake cycles; episodic slip	Dominantly NW-SE-trending fault-parallel (primarily horizontal)
Hydrological	Basin-wide, regional scale	Seasonal and multi-annual trend correlated with precipitation	Primarily vertical and secondary horizontal
Anthropogenic	Localized over range of scales in tens to hundreds of meters	Multi-annual trend and shorter-term seasonal or stochastic variations associated with production practices	Primarily vertical and secondary horizontal

active seismogenic structures (e.g., Chaussard et al., 2015; Lundgren et al., 2009; Lindsey et al., 2014; Tong et al., 2013; Xu et al., 2018). In the recent 2019 Ridgecrest earthquake sequence, an InSAR phase gradient map has elucidated hundreds of linear coseismic strain concentrations or fractures surrounding the rupture (Xu et al., 2020). On the other hand, subsurface water storage, hydrocarbon and geothermal fields have a multitude of dimensions on a smaller scale (e.g., Amelung et al., 1999; Argus et al., 2014; Chaussard & Farr, 2019; Chen et al., 2016; Edwards et al., 2009). Hydrologically driven displacements are distinct between sedimentary basins dominated by confined aquifers and the bounding fault structures (e.g., Chaussard et al., 2014; Hu et al., 2018 and references cited therein). Human activities such as groundwater pumping, hydrocarbon and geothermal extraction, as well as fluid injection, can result in highly localized ground deformation surrounding the clusters of production wells in a field, often exhibiting bull's-eye features in the wrapped interferograms (e.g., Argus et al., 2005; Fielding et al., 1998; Goebel & Shirzaei, 2020; Lanari et al., 2004), and likely trigger earthquakes (e.g., Bawden et al., 2001; Brodsky & Lajoie, 2013; Johnson & Majer, 2017).

Temporarily, all displacement sources involve long-term, multi-annual trends such as from interseismic creep, prolonged climatic fluctuations, and persistent extraction of natural resources (Table 1). Concurrently, the associated displacement time series may also reveal episodic signals due to seismic and aseismic fault slip events, seasonal elastic surface and shallow water loading and poroelastic groundwater volume strain, as well as anthropogenic fluid withdrawal and injection activities subject to weather variations, economic considerations, and workday cycles (e.g., Riel et al., 2018).

Orientationally, tectonic strain accumulation in California is primarily horizontal in the context of the northwest-striking, right-lateral strike-slip SAF system (e.g., Zeng & Shen, 2017, Table 1), but tectonic uplift and subsidence can occur in transpressional and transtensional sections of the plate boundary (e.g., Burgmann et al., 2006; Hammond et al., 2018; Shen & Liu, 2020). Non-tectonic processes mainly deform the Earth vertically (e.g., Chaussard et al., 2021). However, horizontal motions are concentrated in the vicinity of operating wells and along the margins of aquifer basins (e.g., Chaussard et al., 2014; Hoffmann et al., 2003; Xu et al., 2018).

### 2.3. Characterization of Ground Displacements

The variability of spatial, temporal and orientational patterns of different sources provides us an opportunity to sort them out using a sequence of statistical analyses. Spaceborne InSAR geodesy is limited by its one-dimensional LOS displacement measurement. Because of the near-polar orbit trajectory, right-looking LOS, and small to moderate incidence angle (e.g., 30°–46° for Sentinel-1 data), the present-day spaceborne InSAR measurements are most sensitive to vertical motions for both ascending and descending data, are moderately sensitive to east-west motions, and have little sensitivity to north-south motions. Therefore, ascending- and descending-orbit InSAR time-series can be combined to identify ground targets that present resolvable vertical and EW motions.

Hydrologically driven displacements, elastic and poroelastic, are generally correlated with the seasonal precipitation and water level changes. Where the seasonal hydrological deformation is dominated by the vertical component, we find the peak signals in the ascending and descending time series are well correlated. This approach has been proven feasible in Salt Lake Valley, Utah (Hu & Bürgmann, 2020). Considering

that hydroclimatic phenomena in California have a dominant one-year periodicity (e.g., precipitation from Western Regional Climate Center, NOAA's National Climatic Data Center and California Data Exchange Center groundwater level data, and river discharge from U.S. Geological Survey), the attendant ground deformation is dominated by the same annual periodicity. We apply a simple and efficient statistical strategy to determine the presence of seasonal features in the InSAR data and quantify the peak-to-peak amplitude and phase. First, we capture the first-order variation of the non-linear component of the 2015–2019 displacement time series using the superposition of multiple independent sinusoidal functions constrained by the least-squares estimate; we found three sinusoidal functions work best for our data (Hu et al., 2018). This step is in essence a temporal interpolation, but it emphasizes time-dependent variations and allows for changes in amplitude and frequency. Second, we consider the targets to feature seasonality in displacements if the non-linear waveform contains  $2n \pm 2$  peaks plus troughs during an  $n$ -year time span (allows  $\pm 2$  for bias). The longer the time frame, that is, larger  $n$ , the more robust is the determination of seasonality because the  $\pm 2$  bias will be even less significant compared with  $2n$ . Third, we use a one-year period sinusoidal function to determine the best-fit amplitude and phase for the confirmed seasonally deforming targets. Fourth, we remove the seasonal component (if any) from the original time series and estimate the multi-annual rates (mm/yr), which are referred to the geographic location of GPS station LUTZ in the central San Francisco Bay Area (Long.  $-121.87^\circ$ , Lat.  $37.29^\circ$ ; e.g., d'Alessio et al., 2005). In the traditional characterization of InSAR and GNSS time series, seasonal components are typically modeled with sinusoidal functions with periodicity of 1 year (annual) and 0.5 years (semi-annual) along with other model terms, e.g., linear, offsets, postseismic (exponential or logarithm) and higher order polynomial (or other functional forms) for episodic transients. Using the sinusoidal functions of time for time-series InSAR analysis is not new; however, our method determines the seasonality of the targets rather than treating all targets blindly using a sinusoidal function with a fixed annual periodicity.

#### 2.4. Machine Learning - Random Forest

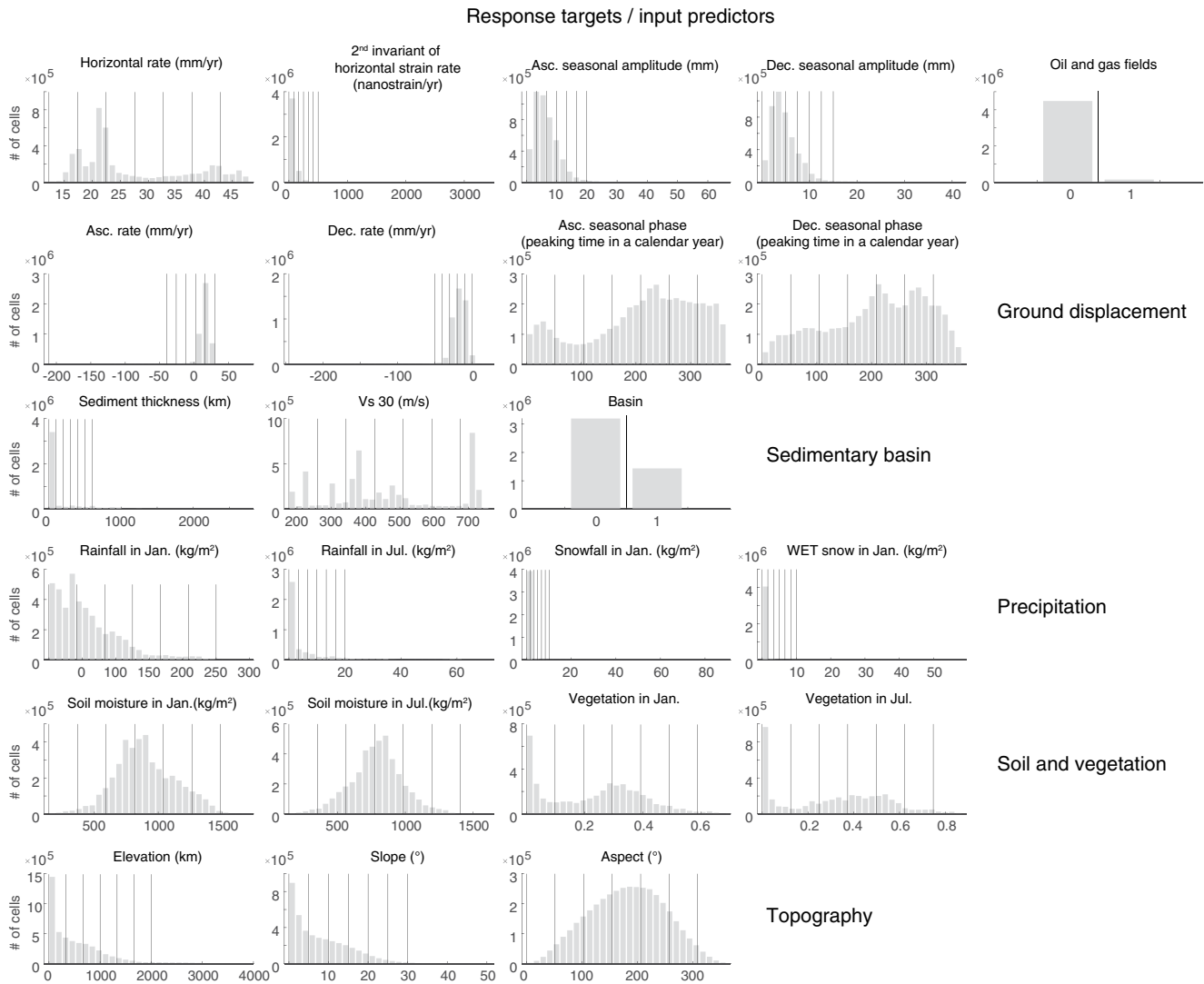
Random Forest (RF) is an ensemble learning algorithm relying on a tree structure (Pal, 2005; Smith & Majumdar, 2020). Each tree splits from the node to proceed to the surrogate trees. The number of trees  $N$  to be grown can be defined by the users. Larger  $N$  and thus a deeper forest enhance the prediction; however, increasing regression trees over 100 does not improve the performance significantly while definitely resulting in higher computation cost (Oshiro et al., 2012; Smith & Majumdar, 2020). A total of  $N$  tree predictors in the forest will vote for an optimal classifier to simulate the response target.

We aim to separate tectonic from hydrological and other non-tectonic processes by comparing their “proxies” (used as the response targets) with the distribution of various “features” (used as the predictors) thought to relate to either type of process. We use the seasonal amplitude and phase in the InSAR-derived time-series displacements and the representative wet and dry months in climatic and soil-related data sets to represent the dominant temporal features in the time-dependent geosystem.

To assemble the features of predictors and response targets for the training datasets, we resample the data sets into the same georeferenced grid cells in the World Geodetic System (WGS84) reference coordinate system. We set the longitude from  $-124.8^\circ$  to  $-114^\circ$  and the latitude from  $31.5^\circ$  to  $41.5^\circ$ . Considering that some data sets have coarse spatial resolutions (e.g.,  $0.125^\circ$  for precipitation- and soil-related maps released by NLDAS) and our state-wide area of interest, we apply an interval of  $0.0025^\circ$  ( $\sim 270$  m) to generate a  $4,001 \times 4,321$  matrix for each feature. We only consider the cells with all observations available, which is in essence determined by the existence of seasonality in ground deformation that was determined from ascending and descending Sentinel-1 time series (Figure S8 in Supporting Information S1). The consequent number of valid cells is 4,633,761. A random 75% of the dataset will be used to establish the RF model for training and the remaining 25% will be used for testing.

The applied features can be fully independent in physical meaning. However, in the numerical representations, these features are divergent in their ranges and orders of magnitude. Therefore, we initialize the categorization in the same levels (i.e., the number of categories) for each feature, except for the groundwater basins and oil and gas fields, for which we only have two categories: within- or outside-of-basin/oil and gas fields. We use fault traces as auxiliary data rather than the input feature to interpret our results.





**Figure 3.** Probability distribution of values for each feature. X-axes show the ranges of the feature values. Y-axes show the number of cells for each sample bin and the total number in every histogram is 4,633,761. The vertical lines indicate the seven equal-sized ranges for most features while some are cut off at their low-density ends of the histograms. The first row represents five response targets for the training sets, which may also function as the predictors for some other training set. The other four rows, considering ground displacement, sedimentary basin, precipitation, soil and vegetation, and topography, represent the common predictors used for all training sets. Only two categories, representing within or outside the feature, are applied for two features—the basins and the oil and gas fields.

Here we have 23 features composing the training sets (Figures 2 and 3; Table S1 in Supporting Information S1). 1–2: tectonic-related horizontal ground displacement rate and second invariant of strain rate from GPS measurements; 3–5: Ascending rate, seasonal amplitude, and seasonal phase (we use the time of year when the ground surface is at peak level measured along the LOS); 6–8: Descending rate, seasonal amplitude, and seasonal phase; 9–11: basin-related data including the geographic locations of basin-fill deposits, the sediment thickness, and shear wave velocity at a depth of 30 m ( $V_s 30$ ); 12–15: precipitation data including rainfall in January and July, snowfall and water equivalent thickness (WET) of snow in January (snow is absent in the summertime and thus we exclude July as the input); 16–19: soil moisture and vegetation fraction in January and July, respectively; 20–22: topographic information including elevation, slope, and aspect; 23: the geographic locations of oil and gas fields.

We consider five training sets. First, to investigate the correlation between various phenomena and the tectonic surface deformation, we choose either (a) the tectonic horizontal velocity field or (b) the derived second invariant of the strain rate as the response targets. Second, to resolve the seasonal deformation mainly

due to cyclic hydrological poroelastic and elastic processes, we use the seasonal displacement amplitude from the (c) ascending and (d) descending Sentinel-1 orbits as individual targets. Ground displacements associated with hydrological processes dominate in the vertical component, and the seasonality is well preserved in the radar LOS measurement after the projection. Third, to investigate the induced phenomena due to human activity, we use (e) the oil and gas fields as the target. All the other data not being used as the response target are taken as input features in each training set. Note that we do not use the features with a direct association to the training set. For example, when (a) the horizontal velocity field is the response target, we do not use (b) the second invariant of strain rate as the predictor; similar scenarios hold for (c) and (d).

We create a regression ensemble by growing a random forest of  $N = 100$  regression trees. A larger number of input features does not guarantee significantly improved performance, as features may correlate and end up with redundancy. On the other hand, increased categorization levels via smaller intervals usually enhance the accuracy but require heavier computation rising nonlinearly. As a start, we group each feature into seven continuous categories with selected cut-off bounds based on their probability distributions (vertical lines in Figure 3), so as to better focus on the high probability values. We grow regression trees and inspect the ensemble performance from the rank of importance, which is computed by dividing the sum of changes in the mean squared error due to splits on every predictor by the number of branch nodes. Larger importance indicates that the predictors have a greater influence on predicting the specified feature, and zero value represents the smallest predictor importance. The predictive measure of the association matrix is another index to quantify the similarity between decision rules that split observations among predictors. Element  $(i, j)$  in the association matrix is the predictive measure of association averaged over surrogate splits on predictor  $j$  for which predictor  $i$  is the optimal split predictor. Larger associations indicate stronger correlated pairs of predictors.

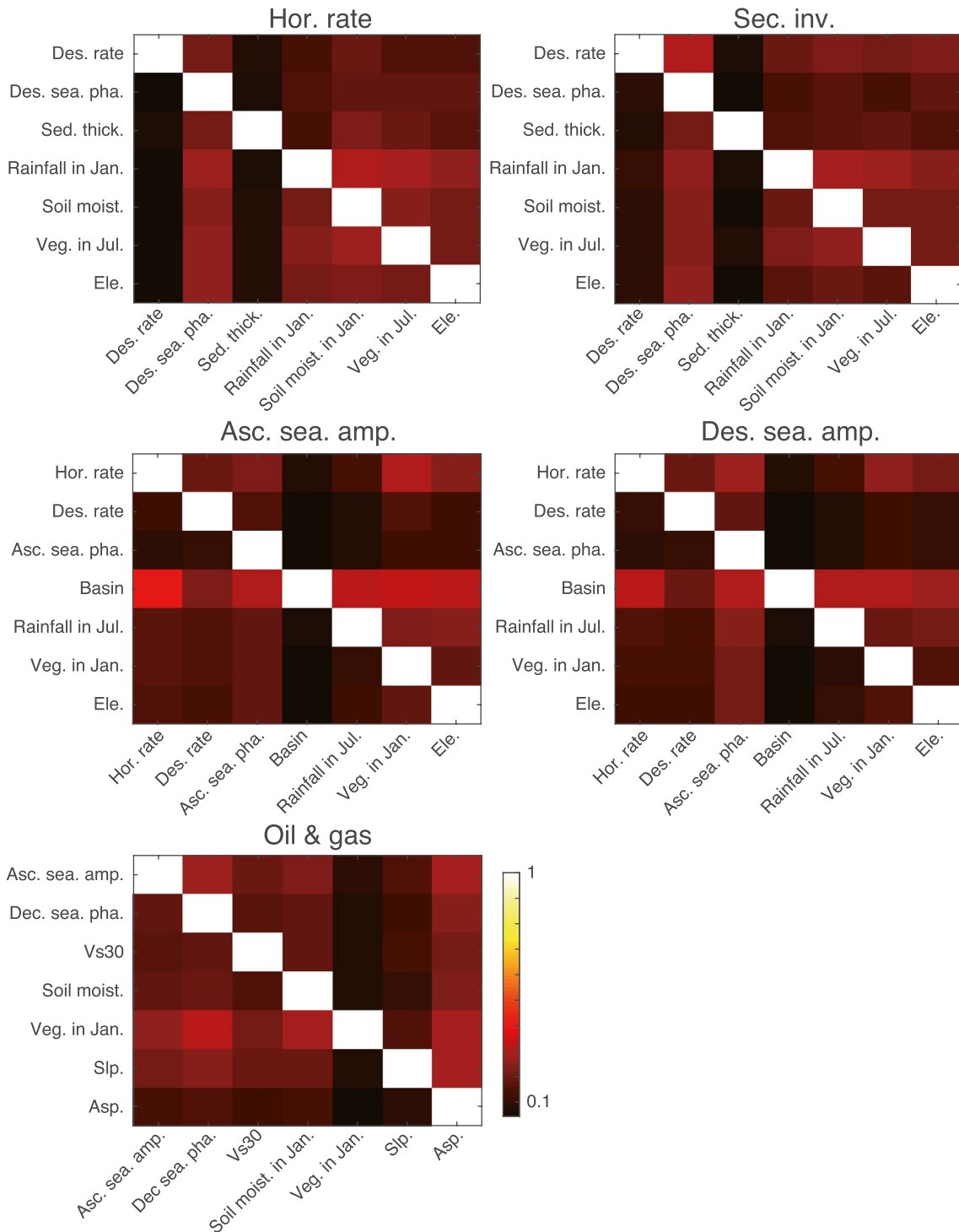
To avoid data redundancy and enhance data diversity, we pick seven most important predictors that do not have strong association between themselves, such that in the training set ii, only the vegetation fraction in July will be considered, even though vegetation fraction in January also ranks high in the importance. Then we increase their levels of categorization to 100 and regrow  $N$  trees for each training set using the same 75% of the data as training dataset. We compute the  $R^2$  between the observation vector  $O$  and the prediction vector  $P$  with the same number of components  $n$ , which is given by  $R^2 = \frac{(\sum_{i=1}^n (O_i - \bar{O})(P_i - \bar{P}))^2}{\sum_{i=1}^n (O_i - \bar{O})^2 \sum_{j=1}^n (P_j - \bar{P})^2}$ , where  $\bar{O} = \sum_{i=1}^n (O_i)/n$  and  $\bar{P} = \sum_{j=1}^n (P_j)/n$ . Here  $R^2$  represents the variability around the mean that can be explained by the regression model. If the new  $R^2$  from the seven tailored predictors is larger than the original  $R^2$  from all available predictors ( $\sim 20$  for each training set), the new model with reduced predictors has a better performance in prediction. Then we predict the responses for the remaining 25% of the dataset using the tuned ensemble of decision trees.

Considering the total feature space  $X$ , partial-dependence plots are often used to visualize the partial dependence of the response to selected predictor variables  $X^s$  by marginalizing over  $X^c$ , which is the complementary set of  $X^s$  in  $X$ . A predicted response  $f(X)$  depends on all variables in  $X$  given by  $f(X) = f(X^s, X^c)$ . The partial dependence of predicted responses on  $X^s$  is equivalent to the expectation of predicted responses on all variables including  $X^c$  (MATHWORKS; <https://www.mathworks.com/help/stats/regressiontree.plot-partialdependence.html>). It helps determine the effect of features  $X^s$  on the prediction and whether the first-order relationship between the target and the feature.

### 3. Results

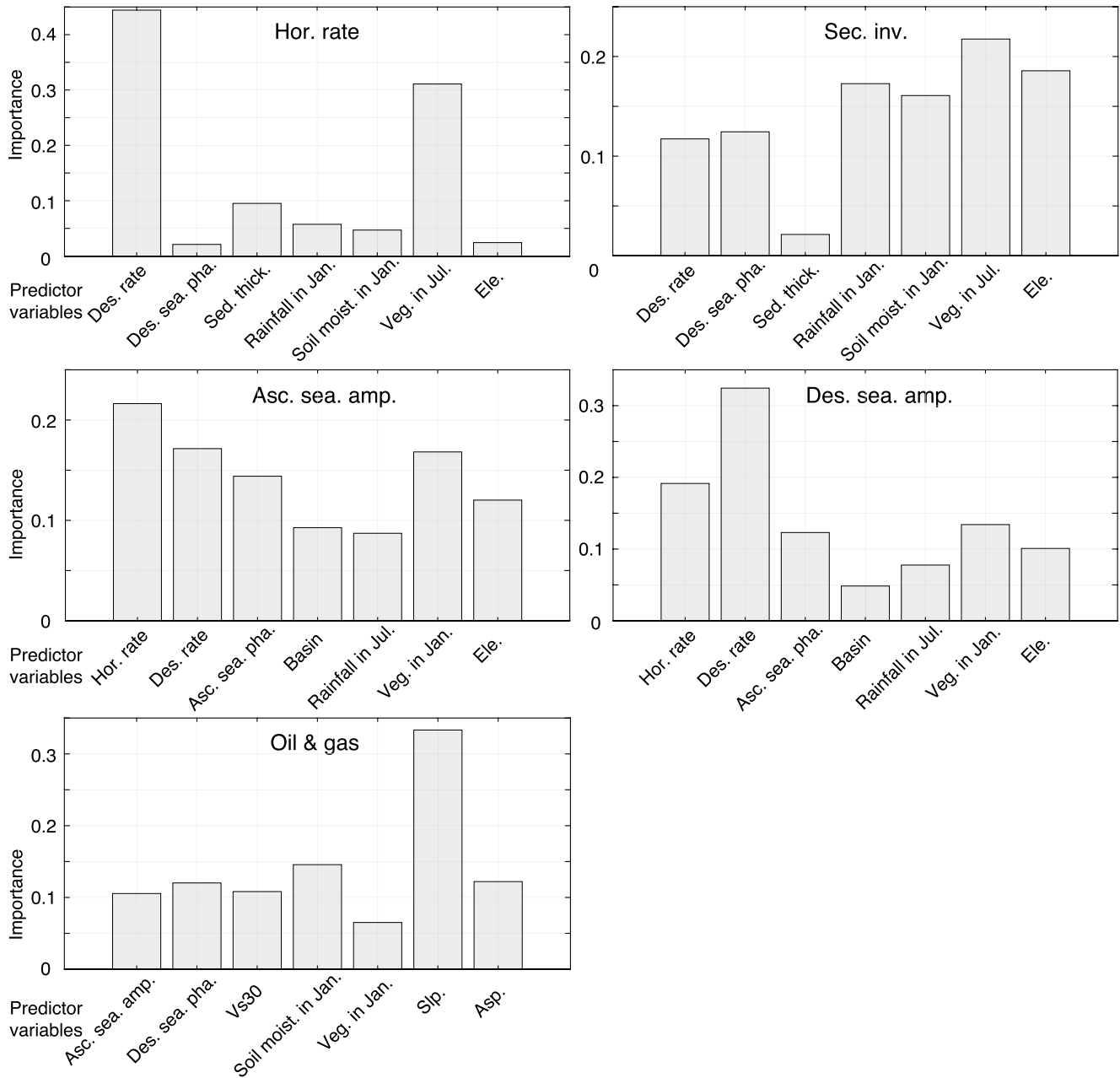
#### 3.1. Predictor Association and Importance Estimates

Predictor association represents the interconnection between pairs of input features. Not surprisingly, especially strong associations exist between two forms of snow data - snowfall and the water equivalent thickness of snow, between snow and elevation, between LOS rates from ascending and descending orbits, and between hydroclimatic features collected in different months (e.g., January and July used for this study) (Figure 4 and Figure S9 in Supporting Information S1). For training sets i and ii, which are designed to predict tectonic horizontal deformation and strain rate, the most important predictor is the radar LOS rates



**Figure 4.** Association between any two features for each training set with the seven most important predictors. The diagonals represent the largest correlation for each feature with itself. Element  $(i, j)$  in the association matrix is the predictive measure of association averaged over surrogate splits on predictor  $j$  for which predictor  $i$  is the optimal split predictor. See Table S1 in Supporting Information S1 for the description of the abbreviations.

(Figure 5 and Figure S10 in Supporting Information S1), which is also not a surprise as they both characterize the long-term surface displacement rate. The descending rate is more important than the ascending rate as the LOS directions of the descending tracks are more aligned with the strike-slip fault traces while the LOS directions of the ascending tracks are more orthogonal to the faults and thus capture less of the



**Figure 5.** The predictor importance estimates in each training set with seven most important predictors. *X* axes show the predictor variables and *Y* axes show the relative importance. See Table S1 in Supporting Information S1 for the description of the abbreviations.

dominant strike-slip deformation signal. Therefore, the descending LOS captures more of the horizontal deformation as a sum of the north and the east vectors. Next, we focus on a tailored group of seven important, while less associated features, including vegetation, descending rate, rainfall, elevation, soil moisture, sediment thickness, and descending seasonal phase, which are generally in a descending order of importance considering their performance in both training sets. However, to be precise, the ranks of importance for training sets with these two individual response targets differ, although the horizontal strain rate is in essence derived from the spatial gradient of the horizontal deformation model and highlight the high-frequency signals.

For training sets iii and iv, which are designed to predict seasonal deformation due to hydrological processes, a tailored group of seven important features include horizontal rate, descending rate, vegetation,

**Table 2**  
Prediction Statistics for Different Training Sets

Training set	Response target	# Of predictors	Predictable $R^2$		Predictor importance		
			Training (75%)	Test (25%)	1	2	3
i	Horizontal displacement rate	21	92.05%	91.94%	Des. rate	Asc. rate	Veg. in Jul.
		7	94.48%	94.23%	Des. rate	Veg. in Jul.	Sed. thick.
ii	Second invariant of strain rate	21	87.36%	85.53%	Ele.	Veg. in Jul.	Veg. in Jan.
		7	92.07%	89.85%	Veg. in Jul.	Hgt.	Rainfall in Jan.
iii	Asc. Seasonal displacement amplitude	21	82.23%	81.15%	Des. rate	Veg. in Jul.	Asc. rate
		7	89.00%	86.00%	Hor. Rate	Des. rate	Veg. in Jan.
iv	Des.	21	79.86%	79.23%	Des. rate	Asc. rate	Veg. in Jan.
		7	90.61%	88.26%	Des. rate	Hor. rate	Veg. in Jan.
v	Oil and gas fields	22	47.84%	44.33%	Slp.	Asp.	Veg. in Jan.
		7	56.99%	34.93%	Slp.	Soil moist. in Jan.	Asp.

*Note.* We consider two scenarios of all and tailored predictors for each training set.  $R^2$  suggests the fraction of the responses that can be predicted by the random forest model. The three most important predictors from all and the tailored predictors are listed.

ascending seasonal phase, elevation, basin, and rainfall, which are generally in a consistently descending order of importance in these two training sets.

The long-term LOS rate is important to characterize both tectonic and hydrological processes. Distinct lineations of high displacement gradient and strain are aligned with the San Andreas Fault system. The aquifers in the Central Valley show large long-term subsidence; for areas with identifiable seasonal motion, the seasonal amplitude is larger than that of the non-aquifer regions. On the other hand, there is no evident contribution from the oil and gas fields to the regional tectonic and hydrological processes (training sets i–iv). Put another way, the prediction of the locations of oil and gas fields in the training set v is not ideal even with a tailored group of seven important features including the slope, soil moisture, aspect, descending seasonal phase, Vs30, ascending seasonal amplitude, and vegetation.

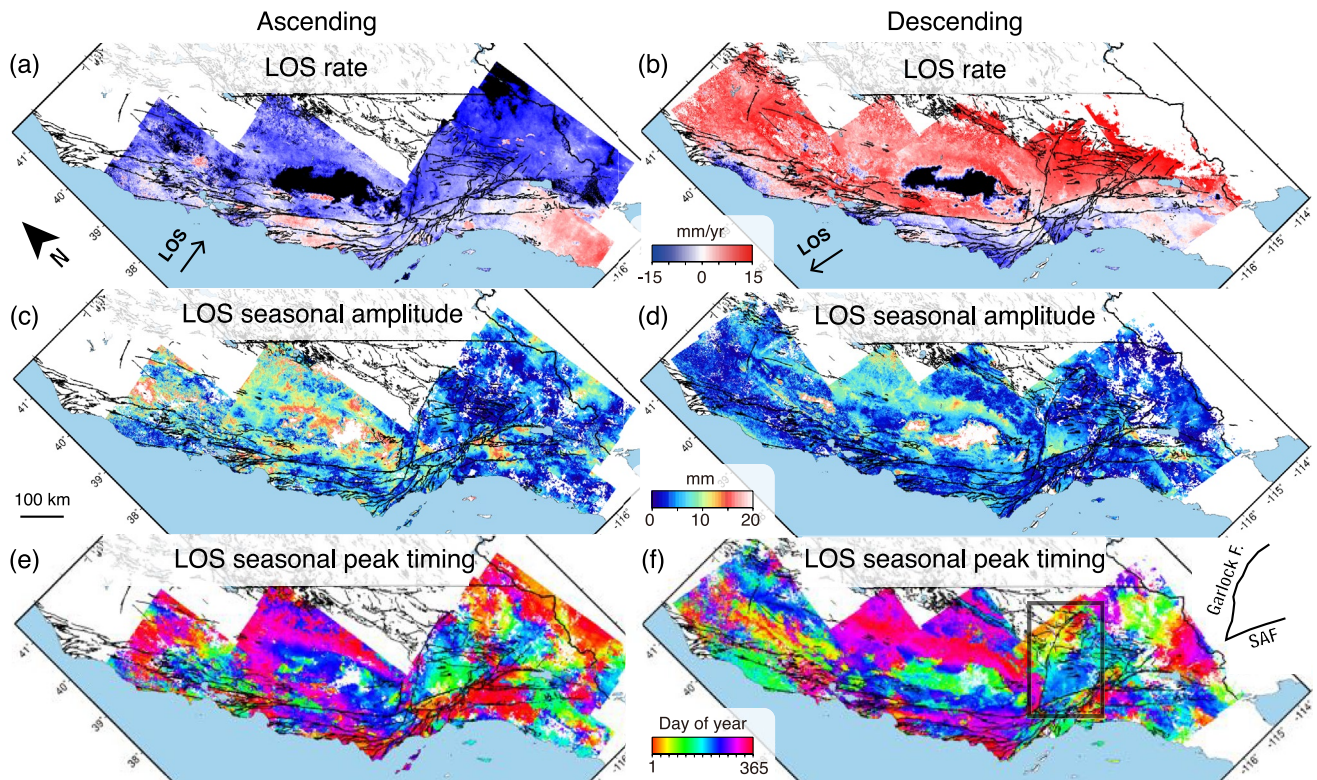
### 3.2. Predictor Performance $R^2$ Estimates

For the training sets i and ii, which are designed for predicting the tectonic deformation, ~95% of the horizontal rate and ~90% of the second invariant of strain rate can be resolved for the training dataset (75% of the total) and test dataset (25% of the total). For the training sets iii and iv, designed for predicting the seasonal displacement amplitude that is most likely associated with hydrological processes, 86%–90% of the data can be resolved (Table 2). The prediction for the oil and gas fields is less promising, for which only 47%–57% of the training dataset can be recovered. Even worse, only ~35% of the test dataset can be explained via the tuned model, suggesting an ineffective estimate.

## 4. Discussions

### 4.1. Long-Term and Seasonal Characteristics of Ground Motions

Independent acquisitions from ascending and descending orbits can be used to cross-validate the spatial distribution of the displacements such as those due to creeping faults and long-term vertical subsidence such as in the Central Valley. Creeping faults represent the near-surface aseismic slip of fault planes driven by ductile shear in the lower crust and the elastic strain in the upper crust. Referring to the position at the plate boundary, the LOS vector of the ascending track points east to north-east while that of the descending track points west to north-west. The Pacific Plate moves toward the satellite in the LOS direction of the ascending orbit while it moves away from the descending satellite; the North American Plate moves away from the ascending satellite while it moves toward the descending satellite. The distinct LOS directions from



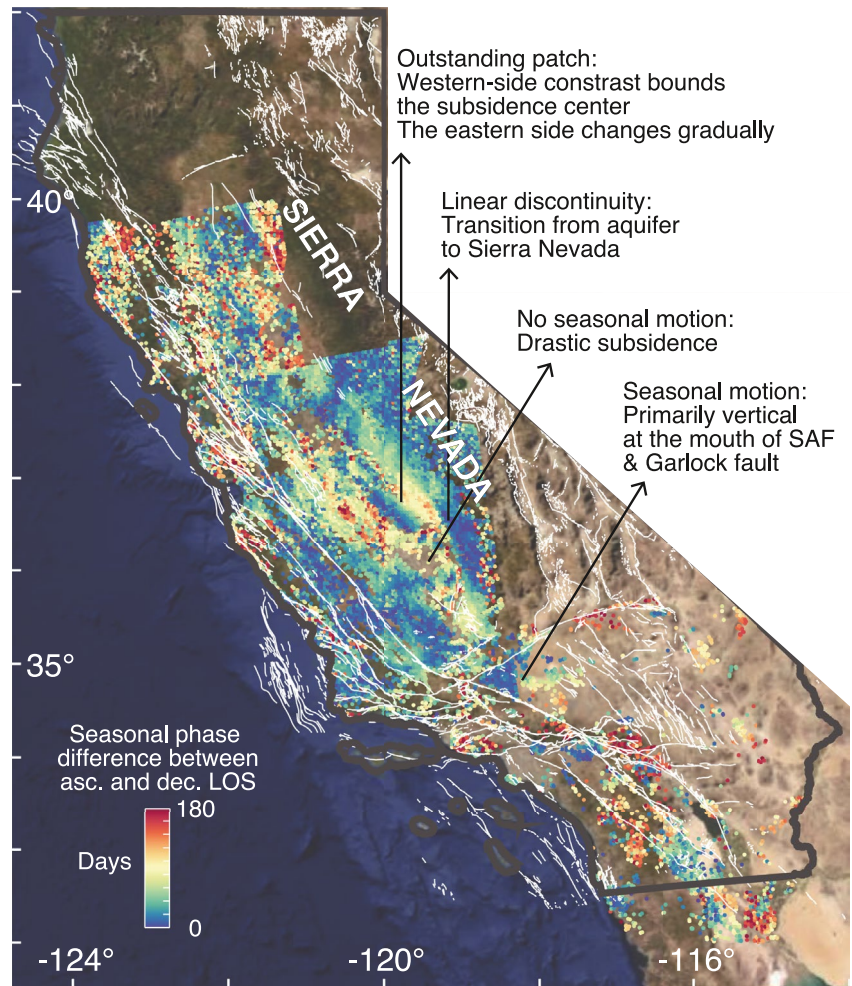
**Figure 6.** Characteristics of time-series displacements measured along Sentinel-1 ascending and descending line-of-sight (LOS). (a) and (b) 2014–2018 LOS velocity (Figure 2 applies larger color scale to reveal the drastic subsidence in the San Joaquin Basin in Central Valley). (c) and (d) LOS peak-to-peak seasonal amplitude. (e) and (f) Time of year of LOS seasonal peak. (a), (c), and (e) are ascending results (T035, T064, T137, and T166). (b), (d), and (f) are descending results (T042, T071, T115, T144, and T173). Black lines show the fault traces. Void (white) area at the center of the Central Valley does not detect seasonal deformation given the predominant rapid subsidence. Boxed area in panel f shows the intersection of Garlock Fault and San Andreas Fault (SAF). See Figure S7 in Supporting Information S1 for an enlarged view with the northside upward.

both satellite orbits consistently demonstrate the right-lateral strike slip at rates of a few to tens of mm/yr along the SAF (Figures 6a and 6b).

The rates and the driving mechanisms vary segmentally along the same or adjacent seismogenic systems in California. Nevertheless, no clear seasonal signal was found in creep records (Turner et al., 2015). The mineral talc in serpentinite rock found along the Hayward fault and central creeping segment of SAF has low shear strength and velocity-strengthening frictional properties, and may be responsible for the shallow aseismic slip observed on these faults (Moore & Rymer, 2007). On the other hand, increased pore pressure in the upper sedimentary layers of the Superstition Hills fault may lead to the creep along this fault (Wei et al., 2009). Alternatively, chemical reactions, dilation of dry rocks, elevated temperatures, and fault geometry may also function as the drivers to creeping faults (Harris, 2017).

The intersection between the Garlock Fault and the San Andreas Fault (inset in Figure 6f) is clearly revealed in the seasonal displacement amplitude and timing, especially evident from the descending tracks. The phase changes across the faults suggest perturbations of groundwater flow and/or differences of hydraulic properties.

Parts of the San Joaquin Basin in the Central Valley have been experiencing inelastic subsidence when the hydraulic head drops below the pre-consolidated condition (Chaussard & Farr, 2019; Liu et al., 2019). Here we find that seasonality in the time-series deformation may exist but is overwhelmed by the very rapid long-term subsidence (shown as no data in the middle of the Central Valley in Figures 6e and 6f and 7). The interbedded layers may have lost their capability to store and release the water in sync with the cyclic weather changes. On the other hand, the areas surrounding the subsidence center show relatively large peak-to-peak seasonal amplitude ( $> \sim 10$  mm), implying poroelastic response of aquifers to seasonal variations



**Figure 7.** The absolute values for the seasonal phase difference in days (within half a year) between the ascending and descending LOS displacement measurements. Here we only show targets with strong seasonal motions (e.g., seasonal amplitude larger than 5 mm).

in precipitation and groundwater. Interestingly, an elongated patch bounding the most rapidly subsiding center in the San Joaquin Basin shows similar timing of seasonal surface peaks in ascending and descending results, unlike the surrounding marginal aquifers with timing discrepancy (Figure 7). This suggests that there is a horizontal component to the seasonal deformation in the marginal areas. The abrupt reduction in subsidence rates and contrasting seasonal phase at the western margins of this patch may suggest a distinct aquifer architecture and groundwater usage; likely representing the boundaries of the overexploited aquifer with irreversible subsidence. The seasonal phase gradients become smooth in individual ascending and descending results, as well as their differences, to the east of this patch. Near the edge of the Mojave Desert, the acute angle formed by the Garlock Fault and SAF is evident in the seasonal amplitude and phase maps, reflecting deformation of the Antelope Valley aquifer (Galloway et al., 1998) (Figure 6). The bedrock southern Sierra Nevada in the InSAR-mapped region shows consistent timing of ground surface peaks in the end of October to early November from ascending and descending data, suggesting that (a) the seasonal motion is mainly vertical and (b) October to November represents the maximum unloading before the arrival of the wet season, consistent with GPS-based results (e.g., Amos et al., 2014) (Figures 6e and 6f and 7).

Localized rapid subsidence inferred from ascending and descending LOS rates  $< -15$  mm/yr (Figures 1, 6a, b, and Figure S7 in Supporting Information S1), smaller in size than that of the inelastic aquifer (e.g., part of the  $>25$ k-square-kilometers San Joaquin Valley) (Chaussard & Farr, 2019), can be found in several principal hydrocarbon and geothermal production fields in Sonoma, Sacramento, Bakersfield, Santa Barbara, and Los

Angeles regions, as well as the Imperial Valley Geothermal Project on the southeastern shore of the Salton Sea (Jiang & Lohman, 2021) (locations labeled in the panel of oil & gas fields in Figure 2). These industrial fields do not present clear evidence of seasonal deformation. The prediction for industrial fields based on the random forest method is not promising for several possible reasons. First, not all oil and gas wells are currently active or will generate appreciable deformation that can be resolved by InSAR. The deforming areas may be too small compared to the standardized grid size ( $0.0025^\circ$ ,  $\sim 270$  m) and thus could be filtered out or smoothed during resampling. In addition, the established production wells can hardly represent all existing sites with natural resources such as hydrocarbon and geothermal energy. With more complete knowledge of the locations of such natural resources and higher resolution of NLDAS products, more reliable prediction and thus planning is possible in the future.

#### 4.2. The Importance of Vegetation Fraction in Predicting Tectonic and Hydrological Processes

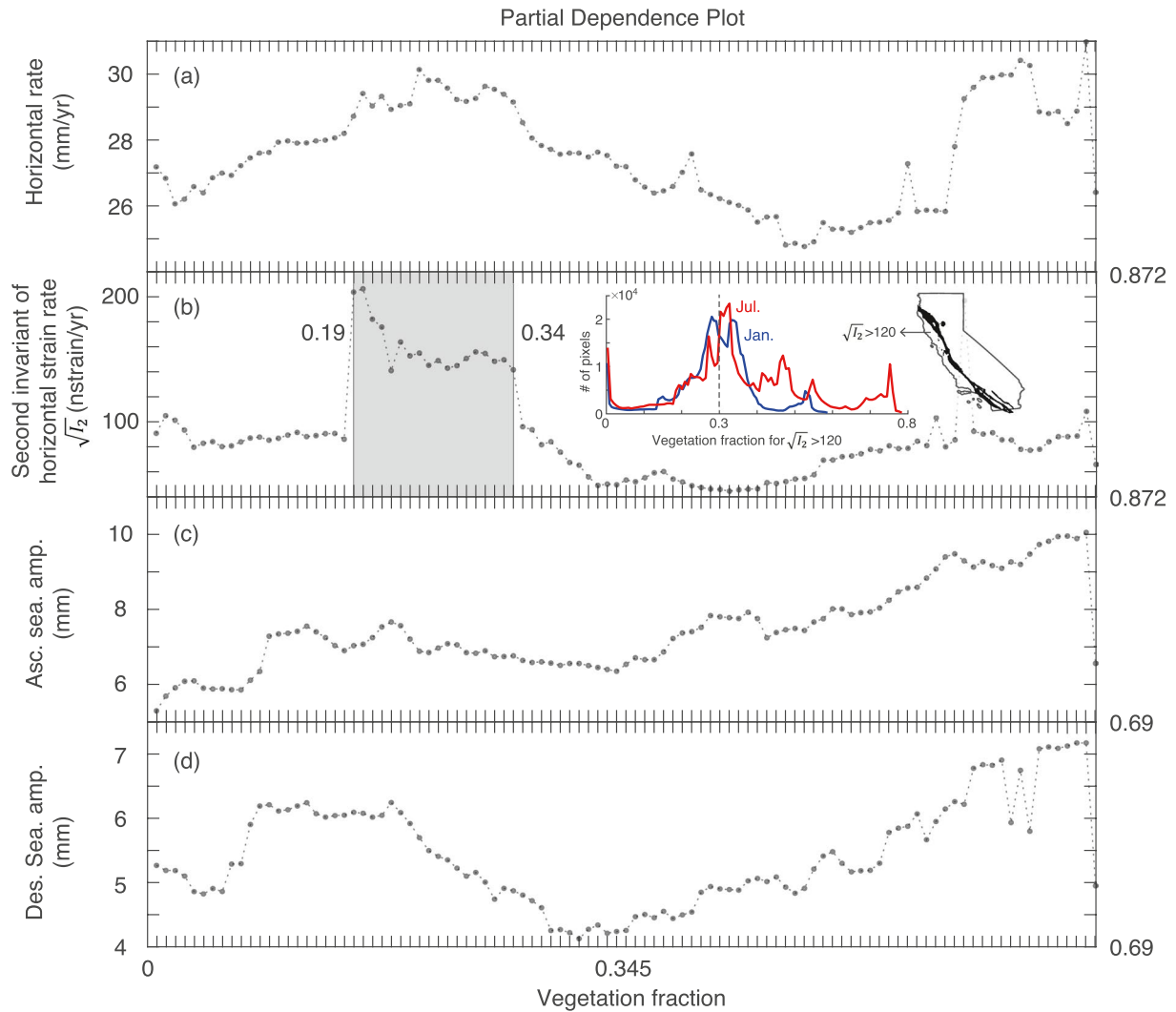
The persistent high importance of the vegetation fraction in predicting the displacement for both hydrological and tectonic processes is remarkable (Figure 5). The partial dependence between the vegetation fraction and the target responses shows that the changes and sometimes inflections occurred through the range of the vegetation fraction (Figure 8). The discontinuities in the tectonic horizontal displacement and strain rate imprint plate boundaries and the historic and potentially future surface-rupture zones, which also correspond to a narrow range of vegetation fraction (0.19–0.34 referring to the condition in 2018 July) comparatively smaller than that of the vegetated covers in the coastal ranges and the agriculture fields in the Central Valley (Figure 9). This relation between the high strain rate belt and low-to-moderate-vegetation alignments may not be a coincidence for the creeping section north-west of Parkfield and the adjoining locked Cholame-Carrizo sections of central SAF. There are some plausible reasons. The rough/high-relief coastal range morphology correlates with active faulting, and the rough/high-relief topography also correlates with the seasonal, orographic rainfall and thus affect the vegetation growth. Besides that, we speculate that the shallow slope and low-relief areas along the high strain rate belt might be constantly modified and thus may challenge widespread root establishment and intense vegetative growth. In comparison, the ranges further from the high-strain fault strands are less disturbed and exhibit higher vegetation fraction. This is unlike the vegetation lineaments due to fault barriers and thus locally enhanced soil moisture. Vegetation lineaments do exist along the San Andreas Fault and Banning Fault in the southern California deserts (Rymer et al., 2002), but the spatial resolution in our study can only resolve patchy changes in vegetation density rather than a narrow vegetation band. On the other hand, the increased rainfall helps forests flourish in northern California, which corresponds to areas of high vegetation and strain rate (Figures 8b and 9).

The partial differential plots of the vegetation fraction in predicting the seasonal displacement amplitude measured by ascending and descending tracks are different. The ascending LOS direction is almost normal to the mostly north-west trending strike-slip faults, and thus captures the least amount of the tectonic motion; in other words, it captures mostly the hydrological motion. Overall, the ascending seasonal amplitude is proportional to the increased vegetation fraction. This is because the ascending seasonal amplitude represents the seasonal variation of subsurface water storage, and vegetation growth reflects the groundwater supply. On the other hand, the descending LOS measurements more strongly reflect the tectonic horizontal motions. There is a pronounced drawdown of the descending seasonal amplitude when the vegetation fraction is between 0.18 and 0.343 (Figure 8d). Hydrologic, magmatic, and faulting interactions around the Long Valley caldera and elsewhere may jointly contribute to the observed multi-annual and seasonal deformation rate changes (Hammond et al., 2019).

#### 4.3. Limitations of Random Forest (RF)

In this study, we use one classic ML approach of RF. The limitation of RF is that the spatial characteristics have not been taken into account as we consider each 2D feature as a 1D array. In the future, we would like to apply the convolutional neural network to the spatially continuous 2D imagery. Another promising direction in this analysis is the selection of computational-cost-efficient features. Here we only consider the representative wet and dry months for climate- and soil-related features, and we apply certain levels of categorization for predictors, which is in essence downsampling. Future avenues include more sophisticated downsampling algorithms, more comprehensive datasets such as monthly or even daily sampled features,



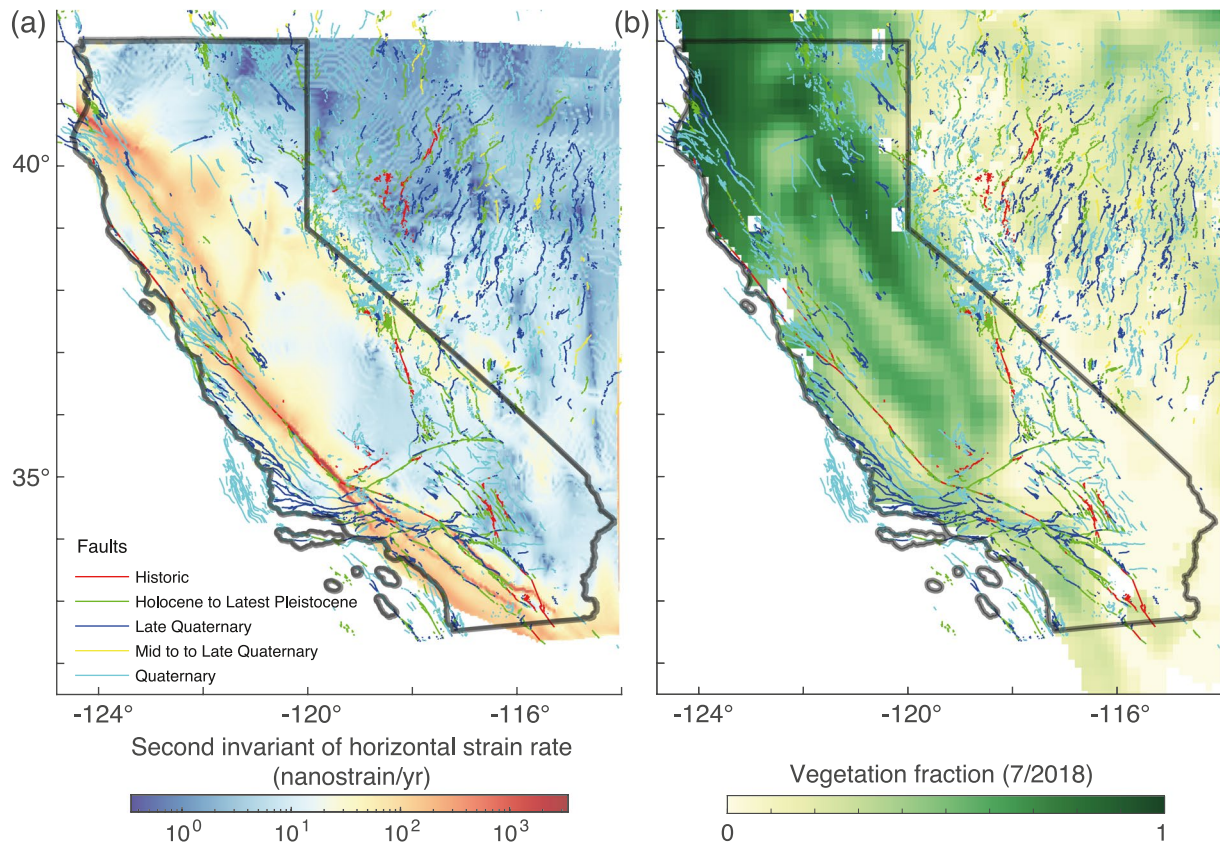


**Figure 8.** Partial dependence plots of the vegetation fraction in predicting different response targets. The shade in panel (b) shows that a comparatively narrow low range of vegetation fractions corresponds to the outstanding large values of the second invariant of horizontal strain rate  $\sqrt{I_2}$ . Inset in panel (b) shows the relation between the vegetation fraction and the number of pixels with second invariant of horizontal strain rate larger than 120 nstrain/yr and their locations are shown to the right. We applied the vegetation fraction in July for training sets (i) and (ii) (panels a and b), and the vegetation fraction in January for training sets (iii) and (iv) (panels c and d) according to the preliminary results (Table 2).

and incorporating principal component analysis or K-mean clustering to extract distinct spatio-temporal patterns, so as to expand the state-wide machine-learning characterization to geodetic time scales.

## 5. Concluding Remarks

This study aims to better characterize the spatiotemporal ground deformation associated with the tectonic, hydrological, and industrial processes in California. Machine learning methods, such as random forest used in this study, are ideal to compile and associate multidisciplinary geodetic, tectonic, topographic, hydrological, and climatic data sets, and to provide insights into the natural complexity and interactions of deformation and surface processes. Here we rely on longitudinally and latitudinally consecutive and partially overlapping paths of ascending and descending Sentinel-1 observations to systematically obtain and decompose the plate-boundary-scale secular and seasonal displacements. We consider the existing long-term horizontal velocities and strain rate field constrained from GPS and tectonic deformation models as proxies for tectonic processes and seasonal displacement amplitude from SAR ascending and descending orbits as the proxies for hydrological process, 86%–95% of which can be predicted by multidisciplinary remote sensing



**Figure 9.** Maps of (a) the second invariant of horizontal strain rate  $\sqrt{I_2}$  and (b) the vegetation fraction (7/2018). A low range of vegetation fraction, high  $\sqrt{I_2}$ , and some of the SAF system overlap in space.

datasets using the random forest algorithm. Surprisingly, vegetation fraction is found to play an important role in the predictions. High-strain-rate zones along the SAF and other faults mainly occur in a narrow range of vegetation fraction ( $\sim 0.3$ ), suggesting that the active faulting, topographic relief, and orographic rainfall may be correlated to affect the vegetation growth.

Linear discontinuities in the long-term, seasonal amplitude and phase of the InSAR displacement fields coincide with some fault strands such as where the SAF and the Garlock fault intersect, suggesting faults partition the groundwater flow. In areas absent of mapped active faults in central California, we identify seasonal phase discontinuities in seasonal displacement at the transition between basin-fill Central Valley and unconsolidated alluvial sediments flanking the Sierra Nevada, suggesting a transition between different seasonal deformation processes involving aquifer volume strain and elastic loading; we also identify a seasonal phase contrast in ascending and descending results at the margins of the inelastically deforming San Joaquin Basin, inferring heterogeneous hydrological units. Some principal oil and gas fields across the California have large displacement rate (e.g.,  $\sim 15$  mm/yr from radar LOS). The oil and gas field proxy is limited as it does not differentiate active from inactive operations and can hardly represent such energy distribution. Our study demonstrates that multidisciplinary datasets from remote sensing and other surveys can be efficiently mined to help better characterize the non-linear processes underlying deformation of the Earth's surface.

### Data Availability Statement

We thank U.S. Geological Survey for providing the thickness of unconsolidated sediments (<https://doi.org/10.5066/P9Z6RC5L>) and NASA for providing the North American Land Data Assimilation System (NL-DAS) Noah Soil Hydraulic Properties Dataset (<https://ldas.gsfc.nasa.gov/nldas/soils>). We also thank the

European Space Agency (ESA) for the extraordinary open data policy of the Sentinel-1 mission and thank the Alaska Satellite Facility (ASF) and UNAVCO for archiving the data and the precise orbital products. The InSAR results were generated with the GMTSAR software developed at the SIO, UC San Diego (<https://topex.ucsd.edu/gmtsar/>). Part of this research was performed at the Jet Propulsion Laboratory, California Institute of Technology under contract with the National Aeronautics and Space Administration, and supported by the Earth Surface and Interior focus area. Figures in this paper were generated using ArcGIS, GMT, and MATLAB.

#### Acknowledgment

This research is supported by the Southern California Earthquake Center (#20026).

#### References

- Amelung, F., Galloway, D. L., Bell, J. W., Zebker, H. A., & Lacznik, R. J. (1999). Sensing the ups and downs of Las Vegas: InSAR reveals structural control of land subsidence and aquifer-system deformation. *Geology*, *27*(6), 483–486. [https://doi.org/10.1130/0091-7613\(1999\)027<0483:stuado>2.3.co;2](https://doi.org/10.1130/0091-7613(1999)027<0483:stuado>2.3.co;2)
- Amos, C. B., Audet, P., Hammond, W. C., Bürgmann, R., Johanson, I. A., & Blewitt, G. (2014). Contemporary uplift and seismicity in central California driven by groundwater depletion. *Nature*, *509*, 483–486. <https://doi.org/10.1038/nature13275>
- Argus, D. F., Fu, Y., & Landerer, F. (2014). GPS as a high resolution technique for evaluating water resources in California. *Geophysical Research Letters*, *41*. <https://doi.org/10.1002/2014gl059570>
- Argus, D. F., Heflin, M. B., Peltzer, G., Crampé, F., & Webb, F. H. (2005). Interseismic strain accumulation and anthropogenic motion in metropolitan Los Angeles. *Journal of Geophysical Research*, *110*, B04401. <https://doi.org/10.1029/2003jb002934>
- Bawden, G. W., Thatcher, W., Stein, R. S., Hudnut, K. W., & Peltzer, G. (2001). Tectonic contraction across Los Angeles after removal of groundwater pumping effects. *Nature*, *412*, 812–815. <https://doi.org/10.1038/35090558>
- Bergen, K. J., Johnson, P. A., De Hoop, M. V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, *363*, eaau0323. <https://doi.org/10.1126/science.aau0323>
- Brodsky, E. E., & Lajoie, L. J. (2013). Anthropogenic seismicity rates and operational parameters at the Salton Sea geothermal field. *Science*, *341*(6145), 543–546. <https://doi.org/10.1126/science.1239213>
- Bürgmann, R., Hilley, G., Ferretti, A., & Novali, F. (2006). Resolving vertical tectonics in the San Francisco Bay area from GPS and Permanent Scatterer InSAR analysis. *Geology*, *34*, 221–224. <https://doi.org/10.1130/g22064.1>
- Carlson, G., Shirzaei, M., Ojha, C., & Werth, S. (2020). Subsidence-derived volumetric strain models for mapping extensional fissures and constraining rock mechanical properties in the San Joaquin Valley, California. *Journal of Geophysical Research*, *125*, e2020JB019980. <https://doi.org/10.1029/2020jb019980>
- Chaussard, E., Bürgmann, R., Fattahi, H., Nadeau, R. M., Taira, T., Johnson, C. W., et al. (2015). Potential for larger earthquakes in the East San Francisco Bay Area due to the direct connection between the Hayward and Calaveras Faults. *Geophysical Research Letters*, *42*, 2734–2741. <https://doi.org/10.1002/2015gl063575>
- Chaussard, E., Bürgmann, R., Shirzaei, M., Fielding, E. J., & Baker, B. (2014). Predictability of hydraulic head changes and characterization of aquifer system and fault properties from InSAR-derived ground deformation. *Journal of Geophysical Research*, *119*, 6572–6590. <https://doi.org/10.1002/2014jb011266>
- Chaussard, E., & Farr, T. G. (2019). A new method for isolating elastic from inelastic deformation in aquifer systems: Application to the San Joaquin Valley, CA. *Geophysical Research Letters*, *46*, 10800–10809. <https://doi.org/10.1029/2019gl084418>
- Chaussard, E., Havazli, E., Fattahi, H., Cabral-Cano, E., & Solano-Rojas, D. (2021). Over a century of sinking in Mexico City: No hope for significant elevation and storage capacity recovery. *Journal of Geophysical Research*, *126*, e2020JB020648. <https://doi.org/10.1029/2020jb020648>
- Chen, J., Knight, R., Zebker, H., & Schreüder, W. (2016). Confined aquifer head measurements and storage properties in the San Luis Valley, Colorado, from spaceborne InSAR observations. *Water Resources Research*, *52*(5), 3623–3636. <https://doi.org/10.1002/2015wr018466>
- d'Alessio, M. A., Johanson, I. A., Bürgmann, R., Schmidt, D. A., & Murray, M. H. (2005). Slicing up the San Francisco Bay Area: Block kinematics and fault slip rates from GPS-derived surface velocities. *Journal of Geophysical Research*, *110*, B06403.
- Dawson, T. E., & Weldon, III, R. J. (2013). *Appendix B: Geologic slip rate data and geologic deformation model, U.S. Geological Survey Open-File Report, 2013-1165-B and California Geological Survey Special Report, 228-B*.
- DeMets, C., Gordon, R. G., & Argus, D. F. (2010). Geologically current plate motions. *Geophysical Journal International*, *181*(1), 1–80. <https://doi.org/10.1111/j.1365-246x.2009.04491.x>
- Diffenbaugh, N. S., Swain, D. L., & Touma, D. (2015). Anthropogenic warming has increased drought risk in California. *Proceedings of the National Academy of Sciences*, *112*, 3931–3936. <https://doi.org/10.1073/pnas.1422385112>
- Dowla, F. U., Taylor, S. R., & Anderson, R. W. (1990). Seismic discrimination with artificial neural networks: Preliminary results with regional spectral data. *Bulletin of the Seismological Society of America*, *80*, 1346–1373.
- Edwards, B. D., Hanson, R. T., Reichard, E. G., & Johnson, T. A. (2009). Characteristics of Southern Californian Coastal aquifer systems. In H. J. Lee, & W. R. Normark (Eds.), *Earth Science in the Urban Ocean: The Southern California Continental Borderland*. (Vol. 454, pp. 319–344). The Geological Society of America Special Paper.
- Evans, E. L. (2018). A comprehensive analysis of geodetic slip-rate estimates and uncertainties in California. *Bulletin of the Seismological Society of America*, *108*(1), 1–18. <https://doi.org/10.1785/0120170159>
- Faunt, C. C. (Ed.). (2009). *Groundwater availability of the Central Valley aquifer, California* (Vol. 1766, pp. 225). U.S. Geological Survey Professional Paper.
- Faunt, C. C., Sneed, M., Traum, J., & Brandt, J. T. (2016). Water availability and land subsidence in the Central Valley, California, USA. *Hydrogeology Journal*, *24*(3), 675–684. <https://doi.org/10.1007/s10040-015-1339-x>
- Fielding, E. J., Blom, R. G., & Goldstein, R. M. (1998). Rapid subsidence over oil fields measured by SAR interferometry. *Geophysical Research Letters*, *25*, 3215–3218. <https://doi.org/10.1029/98gl52260>
- Galloway, D. L., Hudnut, K. W., Ingebritsen, S., Phillips, S. P., Peltzer, G., Rogez, F., & Rosen, P. A. (1998). Detection of aquifer system compaction and land subsidence using interferometric synthetic aperture radar, Antelope Valley, Mojave Desert, California. *Water Resources Research*, *34*(10), 2573–2585. <https://doi.org/10.1029/98wr01285>

- Goebel, T. H. W., Hauksson, E., Aminzadeh, F., & Ampuero, J.-P. (2015). An objective method for the assessment of fluid injection-induced seismicity and application to tectonically active regions in central California. *Journal of Geophysical Research*, *120*, 7013–7032. <https://doi.org/10.1002/2015jb011895>
- Goebel, T. H. W., & Shirzaei, M. (2020). More than 40 years of potentially induced seismicity close to the San Andreas fault in San Ardo, central California. *Seismological Research Letters*, *92*(1), 187–198. <https://doi.org/10.1785/0220200276>
- Hammond, W. C., Blewitt, G., & Kreemer, C. (2016). GPS Imaging of vertical land motion in California and Nevada: Implications for Sierra Nevada uplift. *Journal of Geophysical Research*, *121*, 7681–7703. <https://doi.org/10.1002/2016jb013458>
- Hammond, W. C., Burgette, R. J., Johnson, K. M., & Blewitt, G. (2018). Uplift of the Western Transverse Ranges and Ventura area of Southern California: A four-technique geodetic study combining GPS, InSAR, leveling, and tide gauges. *Journal of Geophysical Research*, *123*, 836–858. <https://doi.org/10.1002/2017jb014499>
- Hammond, W. C., Kreemer, C., Zaliapin, I., & Blewitt, G. (2019). Drought-triggered magmatic inflation, crustal strain and seismicity near the Long Valley Caldera, Central Walker Lane. *Journal of Geophysical Research*, *124*(6), 6072–6091. <https://doi.org/10.1029/2019jb017354>
- Harris, R. A. (2017). Large earthquakes and creeping faults. *Reviews of Geophysics*, *55*(1). <https://doi.org/10.1002/2016rg000539>
- Hoffmann, J., Galloway, D. L., & Zebker, H. A. (2003). Inverse modeling of interbed storage parameters using land subsidence observations, Antelope Valley, California. *Water Resources Research*, *39*(2), 1031. <https://doi.org/10.1029/2001wr001252>
- Hough, S. E., & Bilham, R. (2018). Poroelastic stress changes associated with primary oil production in the Los Angeles Basin, California. *The Leading Edge*, *37*, 108–116. <https://doi.org/10.1190/tle37020108.1>
- Hough, S. E., & Page, M. T. (2016). Potentially induced earthquakes during the early twentieth century in the Los Angeles Basin. *Bulletin of the Seismological Society of America*, *106*(6), 2419–2435. <https://doi.org/10.1785/0120160157>
- Hsu, Y.-J., Fu, Y., Bürgmann, R., Hsu, S.-Y., Lin, C.-C., Tang, C.-H., & Wu, Y.-M. (2020). Assessing seasonal and interannual water storage variations in Taiwan using geodetic and hydrological data. *Earth and Planetary Science Letters*, *550*, 116532. <https://doi.org/10.1016/j.epsl.2020.116532>
- Hu, X., & Bürgmann, R. (2020). Aquifer dynamics and implications to seismic hazard in metropolitan Salt Lake Valley. *Earth and Planetary Science Letters*, *547*, 116471. <https://doi.org/10.1016/j.epsl.2020.116471>
- Hu, X., Lu, Z., & Wang, T. (2018). Characterization of hydrogeological properties in Salt Lake Valley, Utah using InSAR. *Journal of Geophysical Research*, *123*. <https://doi.org/10.1029/2017jg004497>
- Jiang, J., & Lohman, R. B. (2021). Coherence-guided InSAR deformation analysis in the presence of ongoing land surface changes in the Imperial Valley, California. *Remote Sensing of Environment*, *253*, 112160. <https://doi.org/10.1016/j.rse.2020.112160>
- Johnson, C. W., Fu, Y., & Bürgmann, R. (2017). Seasonal water storage, stress modulation, and California seismicity. *Science*, *356*(6343), 1161–1164. <https://doi.org/10.1126/science.aak9547>
- Johnson, C. W., Totten, E. J., & Bürgmann, R. (2016). Depth migration of seasonally induced seismicity at The Geysers geothermal field. *Geophysical Research Letters*, *43*. <https://doi.org/10.1002/2016gl069546>
- Johnson, L. R., & Majer, E. L. (2017). Induced and triggered earthquakes at the Geysers geothermal reservoir. *Geophysical Journal International*, *209*, 1221–1238. <https://doi.org/10.1093/gji/ggx082>
- Kreemer, C., & Hammond, W. C. (2007). Geodetic constraints on areal changes in the Pacific–North America plate boundary zone: What controls Basin and Range extension? *Geology*, *35*(10), 943–946. <https://doi.org/10.1130/g23868a.1>
- Kreemer, C., & Zaliapin, I. (2018). Spatiotemporal correlation between seasonal variations in seismicity and horizontal dilatational strain in California. *Geophysical Research Letters*, *45*, 9559–9568. <https://doi.org/10.1029/2018gl079536>
- Lanari, R., Lundgren, P., Manzo, M., & Casu, F. (2004). Satellite radar interferometry time series analysis of surface deformation for Los Angeles, California. *Geophysical Research Letters*, *31*, L23613. <https://doi.org/10.1029/2004gl021294>
- Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*, 3–10. <https://doi.org/10.1016/j.gsf.2015.07.003>
- Lindsey, E. O., & Fialko, Y. (2013). Geodetic slip rates in the southern San Andreas Fault system: Effects of elastic heterogeneity and fault geometry. *Journal of Geophysical Research: Solid Earth*, *118*, 689–697. <https://doi.org/10.1029/2012jb009358>
- Lindsey, E. O., Fialko, Y., Bock, Y., Sandwell, D. T., & Bilham, R. (2014). Localized and distributed creep along the southern San Andreas Fault. *Journal of Geophysical Research: Solid Earth*, *119*, 7909–7922. <https://doi.org/10.1002/2014jb011275>
- Liu, Z., Liu, P. W., Massoud, E., Farr, T. G., Lundgren, P., & Famiglietti, J. S. (2019). Monitoring groundwater change in California's Central Valley using Sentinel-1 and GRACE observations. *Geosciences*, *9*, 436. <https://doi.org/10.3390/geosciences9100436>
- Lundgren, P., Hetland, E. A., Liu, Z., & Fielding, E. J. (2009). Southern San Andreas-San Jacinto fault system slip rates estimated from earthquake cycle models constrained by GPS and interferometric synthetic aperture radar observations. *Journal of Geophysical Research*, *114*(B2). <https://doi.org/10.1029/2008jb005996>
- McPhillips, D. F., Herrick, J. A., Ahdi, S., Yong, A. K., & Haefner, S. (2020). *Updated compilation of VS30 data for the United States*. U.S. Geological Survey data release. <https://doi.org/10.5066/P9H5QEAC>
- Miller, J. A. (2000). *Groundwater Atlas of the United States*. U.S. Geological Survey Hydrologic Investigations Atlas 730-D.
- Molnar, P., & Dayem, K. E. (2010). Major intracontinental strike-slip faults and contrasts in lithospheric strength. *Geosphere*, *6*(4), 444–467. <https://doi.org/10.1130/ges00519.1>
- Moore, D. E., & Rymer, M. J. (2007). Talc-bearing serpentinite and the creeping section of the San Andreas fault. *Nature*, *448*, 795–797. <https://doi.org/10.1038/nature06064>
- Ojha, C., Shirzaei, M., Werth, S., Argus, D. F., & Farr, T. G. (2018). Sustained groundwater loss in California's Central Valley exacerbated by intense drought periods. *Water Resources Research*, *54*, 4449–4460. <https://doi.org/10.1029/2017wr022250>
- Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). *How many trees in a random forest?* Machine Learning and Data Mining in Pattern Recognition, 154–168. [https://doi.org/10.1007/978-3-642-31537-4\\_13](https://doi.org/10.1007/978-3-642-31537-4_13)
- Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. <https://doi.org/10.1080/01431160412331269698>
- Poland, J. F., Ireland, R., Lofgren, B., & Pugh, R. (1975). *Land subsidence in the San Joaquin Valley, California, as of 1972*. U.S. Geological Survey Professional Paper 437-H.
- Riel, B., Simons, M., Ponti, D., Agram, P., & Jolivet, R. (2018). Quantifying ground deformation in the Los Angeles and Santa Ana Coastal Basins due to groundwater withdrawal. *Water Resources Research*, *54*. <https://doi.org/10.1029/2017wr021978>
- Ross, Z., Idini, B., Jia, Z., Stephenson, O. L., Zhong, M. Y., Wang, X., et al. (2019). Hierarchical interlocked orthogonal faulting in the 2019 Ridgecrest earthquake sequence. *Science*, *366*(6463), 346–351. <https://doi.org/10.1126/science.aaz0109>

- Rymer, M. J., Seitz, G. G., Weaver, K. D., Orgil, A., Faneros, G., Hamilton, J. C., & Goetz, C. (2002). Geologic and paleoseismic study of the Lavic Lake Fault at Lavic Lake Playa, Mojave Desert, Southern California. *Bulletin of the Seismological Society of America*, 92(4), 1577–1591. <https://doi.org/10.1785/0120000936>
- Shah, A. K., & Boyd, O. S. (2018). *Depth to basement and thickness of unconsolidated sediments for the western United States—Initial estimates for layers of the U.S. Geological Survey National Crustal Model. (No. 2018-1115)*. U.S. Geological Survey.
- Shen, Z.-K., & Liu, Z. (2020). Integration of GPS and InSAR data for resolving 3-dimensional crustal deformation. *Earth and Space Science*, 7, e2019EA001036. <https://doi.org/10.1029/2019ea001036>
- Shen, Z.-K., Wang, M., Zeng, Y., & Wang, F. (2015). Optimal Interpolation of spatially discretized geodetic data. *Bulletin of the Seismological Society of America*, 105(4), 2117–2127. <https://doi.org/10.1785/0120140247>
- Small, P., Gill, D., Maechling, P. J., Taborda, R., Callaghan, S., Jordan, T. H., et al. (2017). The SCEC Unified Community velocity model software framework. *Seismological Research Letters*, 88(5), 1539–1552. <https://doi.org/10.1785/0220170082>
- Smith, R. G., Knight, R., Chen, J., Reeves, J. A., Zebker, H. A., Farr, T., et al. (2017). Estimating the permanent loss of groundwater storage in the southern San Joaquin Valley, California. *Water Resources Research*, 53, 2133–2148. <https://doi.org/10.1002/2016wr019861>
- Smith, R. G., & Majumdar, S. (2020). Groundwater storage loss associated with land subsidence in western US mapped using machine learning. *Water Resources Research*, e2019WR026621.
- Tong, X., Sandwell, D. T., & Smith-Konter, B. (2013). High-resolution interseismic velocity data along the San Andreas Fault from GPS and InSAR. *Journal of Geophysical Research*, 118, 369–389. <https://doi.org/10.1029/2012jb009442>
- Trugman, D. T., Shearer, P. M., Borsa, A. A., & Fialko, Y. (2016). A comparison of long-term changes in seismicity at The Geysers, Salton Sea, and Coso geothermal fields. *Journal of Geophysical Research*, 121, 225–247. <https://doi.org/10.1002/2015jb012510>
- Turner, R. C., Shirzaei, M., Nadeau, R. M., & Bürgmann, R. (2015). Slow and Go: Pulsing slip rates on the creeping section of the San Andreas Fault. *Journal of Geophysical Research*, 120. <https://doi.org/10.1002/2015jb011998>
- Wei, M., Sandwell, D., & Fialko, Y. (2009). A silent M4.8 event of October 3–6, 2006, on the Superstition Hills Fault, Southern California. *Journal of Geophysical Research*, 114, B07402. <https://doi.org/10.1029/2008jb006135>
- Xu, W., Wu, S., Materna, K., Nadeau, R., Floyd, M., Funning, G., et al. (2018). Interseismic ground deformation and fault slip rates in the greater San Francisco Bay Area from two decades of space geodetic data. *Journal of Geophysical Research*, 123, 8095–8109. <https://doi.org/10.1029/2018jb016004>
- Xu, X., Sandwell, D. T., Klein, E., & Bock, Y. (2021). Integrated Sentinel-1 InSAR and GNSS time-series along the San Andreas fault system. *Earth and Space Science Open Archive*. <https://www.essoar.org/doi/abs/10.1002/essoar.10507566.1>
- Xu, X., Sandwell, D. T., Ward, L. A., Milliner, C. W. D., Smith-Konter, B. R., Fang, P., & Bock, Y. (2020). Surface deformation associated with fractures near the 2019 Ridgecrest earthquake sequence. *Science*, 370(6516), 605–608. <https://doi.org/10.1126/science.abd1690>
- Zeng, Y., & Shen, Z.-K. (2017). A fault-based model for crustal deformation in the Western United States based on a combined inversion of GPS and geologic inputs. *Bulletin of the Seismological Society of America*, 107(6), 2597–2612. <https://doi.org/10.1785/0120150362>