north of the equator, based on information from ringing and field data [40, 49]. Because birds were thereafter kept at a simulated latitude of 10°N for further study, the end of migratory restlessness, gonadal regression, and post-nuptial moult were not analyzed.

Birds were weighed and checked for moult at least weekly, and every 2-3 days during the post-juvenile period, by the Institute team led by the authors of the original flycatcher study [13]. We checked body moult by inspecting the entire bird and scored presence of moult if we detected feather growth in any of 19 defined body areas. Wing moult was scored for each flight feather of the right wing following [50]. In addition, starting in their first winter, birds of both sexes were assessed for the state of their reproductive development (testis diameter in males, diameter of the largest follicle in females) by laparotomy approximately every three weeks [51].

To quantify the timing of migratory restlessness, we measured activity continuously to identify phases of nocturnal activity [42]. Activity was recorded throughout the study period via microswitches attached to the perches. We then derived the number of 30-min intervals showing any activity during the night (i.e., during the lights-off period, discounting immediate effects of switching on and off of the lights). We analyzed the resulting time series of nocturnal activity with a changepoint algorithm that defines the start and end of migratory restlessness [42].

Because the point of our experiment was to investigate whether flycatchers had changed their behavior compared to the original captivity experiment 21 years ago, we took particular care to ascertain that in 2002 we quantified the birds' behavior in the same ways as in 1981, and that no systematic measurement bias occurred between replicates. In 2002, microswitch data were collected electronically for all birds by computer-based event recorders. In 1981, the microswitches were attached to an inkwriter (Esterline Angus, Washington USA). The inkwriter recorded activity onto time-charted paper rolls, after which the ink marks were hand-counted by an observer. For each 30 min interval on the recording paper that showed an ink mark during night hours, a bird was scored as "active" for that interval.

In order to minimize differences between the 1981 and 2002 replicates, we carried out two calibration steps of recording methods. The first involved comparing activity recording by inkwriters to those of electronic event recorders. In 2002, in parallel to electronic event recorders, we recorded activity with two Esterline-Angus inkwriters from the original stock, which we moved between cages during the entire recording period. In each cage, birds were recorded simultaneously by both methods for one week, and then the inkwriters were moved to the next bird, so that 2-3 weeks of comparative data were available for all birds. We then hand-counted the ink recordings for comparison with the parallel electronic recordings. Using a linear mixed-effects model (n = 328 nights of paired recordings), we quantified the methods' repeatability and the mean difference between them: repeatability was high (0.951), and the mean difference was 0.75 (95% CI 0.56 to 0.95).

Additionally, we calibrated our hand-counting in 2002 against hand-counting in 1981 using the original ink paper rolls of 5 birds from the 1981 experiment. Our new counts were compared against those noted in the original scoring sheets from 1981 for the same birds. The repeatability (quantified as above) was 0.952 (n = 590 recounted nights). The recounting slightly overestimated activity compared to the original count (mean = 1.01, 95% CI 0.93 to 1.10).

Thus, the calibration data indicated close correspondence between the methods. The slight deviations in both steps are expected to partially offset each other. The original observer of ink counts had counted somewhat more conservatively, but the new electronic method, in turn, was slightly more conservative than the inkwriter. Remaining small mean differences between methods were not expected to affect outcomes because we generated bird-specific estimates for start and end of migratory restlessness by changepoint analysis, which uses relative differences in time series [42]. Thus, we are confident that we measured behavior equivalently in the two replicates.

Description of field study

We obtained field information from three sites, which, like the origin of the captive population, were all located in the Upper Rhine valley (Figure S1; see there for distances). One site is an active study location of free-living flycatchers [34]. The remaining two sites are weather stations, which framed the flycatcher sites to the north and south within the Rhine valley.

Breeding phenology

To assess changes in local pied flycatcher breeding phenology in the wild during the study period, we used a 46-year dataset from Harthausen near Speyer, Germany (49.3° N / 8.4° E; elevation 105 m asl; Figure S1). From 1973-2018, authors DH and UH collected information on the timing of clutch initiation (laydate), hatching, and breeding success of a population of flycatchers, monitoring 55 ± 14 nests per year, of which we obtained laydate information from 40 ± 15 per year. Data were gathered as part of a ringing study in a nest-box population situated in a mixed coniferous/deciduous woodland at 100 m asl [34]. First arrival of birds was in the first ten days of April (range: 1 to 9 April; data from 15 years). Mean clutch size was 6 eggs, mean incubation period 12 days (12.4 ± 1.73 days; n = 49 nests from 2 years; Hoffmann, unpubl.), and on rare occasions birds were double-brooded. To focus on changes at the start of the breeding season, we followed [16] by only including clutches initiated within 30 days of the mean laydate of the first five nests in a given year. In total, we analyzed laydates from 1,834 clutches over 46 years (998 of which occurred in the 21 years spanning the captive studies). In our phenology analyses, we used the mean laydate for each year.

Local ambient spring temperature

We obtained local hourly ambient temperature data from two weather stations in southwest Germany (German weather service, https://www.dwd.de/EN/climate_environment/cdc/cdc_node.html; Figure S1): Mannheim (station ID 5906; 49.47°N, 8.50°E) and Freiburg (station ID 1443; 48.02°N, 7.83°E) from 1973 to 2018. Ambient temperatures of the two stations were closely correlated during the study period (r = 0.96). We averaged the temperature data from these two stations to develop a single regional temperature